Towards Open-Universe Image Parsing with Broad Coverage

Joseph Tighe
University of North Carolina
at Chapel Hill
jtighe@cs.unc.edu

Svetlana Lazebnik
University of Illinois
at Urbana-Champaign
slazebni@illinois.edu

This paper presents an overview of our work on image parsing, which we define as the problem of labeling each pixel in an image with its semantic category. Our aim is to achieve broad coverage across hundreds of object categories, many of them sparsely sampled. We first describe our baseline nonparametric region-based parsing system. This approach is based on lazy learning, and it can easily scale to datasets with tens of thousands of images and hundreds of labels. We then present three extensions to this baseline system. First, we simultaneously label each region as a semantic class (e.g., tree, building, car) and geometric class (sky, vertical, ground) while enforcing coherence between the two label types (roads can't be labeled as vertical). Second, we extend this simultaneous labeling to an arbitrary number of label types. For example, we may want to simultaneously label every image region according to its basic-level object category (car, building, road, tree, etc.), superordinate category (animal, vehicle, manmade object, natural object, etc.), geometric orientation (horizontal, vertical, etc.), and material (metal, glass, wood, etc.). Finally, we present a hybrid parsing system that combines our region-based system with per-exemplar sliding window detectors to improve parsing performance on small object classes, giving broader coverage.

1 Introduction

Our work addresses the problem of image parsing, or labeling each pixel in an image with its semantic category. Many approaches to this problem have been proposed recently, including ones that estimate labels pixel by pixel [14, 19, 32], ones that aggregate features over segmentation regions [10, 17, 23, 27], and ones that predict object bounding boxes [3, 7, 15, 30]. Most of these methods operate in a *closed-universe* scenario, i.e., on datasets with a fixed number of images and a few pre-defined classes. This scenario requires a generative or discriminative model to be trained offline for each class. Training can take days and must be repeated from scratch if new training examples or new classes are added to the dataset.

By constrast, our work focuses on *open-universe* datasets that do not have pre-defined set of labels and can evolve over time to include new images or labels. Recently, a few researchers have begun advocating nonparametric, data-driven approaches suitable for open-universe datasets [37, 22, 21]. Such approaches do not do any training at all. Instead, for each new test image, they try to retrieve the most similar training images and transfer the desired information from the training images to the query. Liu et al. [21] have proposed a nonparametric label transfer method based on estimating "SIFT flow," or a dense deformation field between images. The biggest drawback of this method is that the optimization problem for finding the SIFT flow is fairly complex and expensive to solve. Moreover, the formulation of scene matching in terms of estimating a dense per-pixel flow field is not necessarily in accord with our intuitive understanding of scenes as collections of discrete objects defined by their spatial support and class identity.

We set out to implement a nonparametric solution to image parsing that is as straightforward and efficient as possible, and that relies only on operations that can easily scale to ever larger image collections and sets of labels. Figure 1 gives an overview of our initial region-based parsing system, which will be described in more detail in Section 2. Similarly to [21], our method makes use of a *retrieval set* of scenes whose content is used to interpret the test image. However, unlike the approach of [21], which works best if the retrieval set images are very similar to the test image in terms of spatial layout of the classes, we transfer labels at the level of *superpixels* [28], or coherent image regions produced by a bottom-up segmentation method. The label transfer is accomplished with a fast and simple nearest-neighbor search algorithm, and it allows for more variation between the layout of the test image and the images in the retrieval set. Moreover, using segmentation regions as a unit of label transfer gives better spatial support for aggregating features that could belong to the same object [11].

Our non-parametric parsing system requires very little training and is able to scale to datasets of thousands of images, while still outperforming the parsing performance of [21], which at the time was the state-of-the-art system. Having introduced our baseline parsing system in Section 2, the remainder of the paper will discuss a number of extensions to improve its performance.

First, we leverage geometric/semantic context in the manner of Gould et al. [10]. Namely, for each superpixel in the image, we simultaneously estimate a *semantic* label (e.g., building, car, person, etc.) and a *geometric* label (sky, ground, or vertical surface) while making sure the two types of labels assigned to the same region are consistent (e.g., a building has to be vertical, road horizontal, and so on). Our experiments show that enforcing this coherence improves the performance of both labeling tasks.

Next, we generalize this notion of geometric/semantic labeling to an arbitrary number of label types. The question is what type of labeling to use. We can label image regions with basic-level category names such as grass, sheep, cat, and person; as well as, coarser superordinate-level labels such as animal, vehicle, manmade object, natural object, etc. We can assign geometric labels such as horizontal, vertical and sky, as discussed above. We can also assign material labels such as skin, metal, wood, glass, etc. Further,

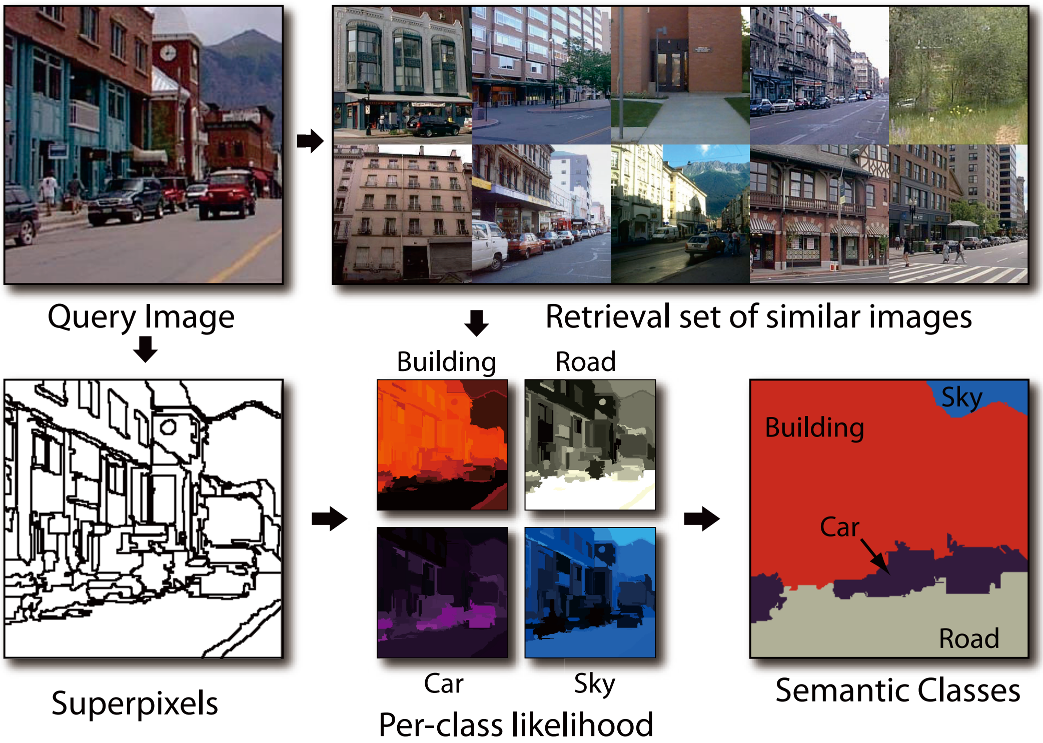

Figure 1. Overview of our region-based image parsing system. Given a query image, we retrieve similar images from our dataset using several global features. Next, we divide the query into superpixels and compute a per-superpixel likelihood ratio score for each class based on nearest-neighbor superpixel matches from the retrieval set. These scores, in combination with a contextual MRF model, yield a final labeling of the query image.

some regions belonging to structured, composite objects may be given labels according to their part identity: if a region belongs to a car, it may be a windshield, a wheel, a side door, and so on.

Here the goal is to understand scenes on multiple levels: rather than assigning a single label to each region, we wish to assign multiple labels simultaneously, such as a basic-level category name, a superordinate category name, material, and part identity. By inferring all the labellings jointly we can take into account constraints of the form "roads are horizontal," "cars are made of metal," "cars have wheels" but "horses have legs," leading to an improved interpretation of the image. Our methods for exploiting geometric/semantic context and performing more general multi-level inference will be described in Section 3.

Finally, Section 4 will discuss our most recent work on incorporating object detectors into our parsing system with the goal of achieving *broad coverage* – the ability to recognize hundreds or thousands of object classes that commonly occur in everyday street scenes and indoor environments. A major challenge in doing this is posed by the non-uniform statistics of these classes in realistic scene images. A small number of classes – mainly ones associated with large regions or "stuff," such as road, sky, trees, buildings, etc. – constitute the majority of all image pixels and object instances in the dataset.

"Stuff" categories have no consistent shape but fairly consistent texture, so they can be adequately handled by image parsing systems based on pixel- or region-level features [4, 5, 21, 25, 32, 35] which includes our base parsing system.However, these systems have difficulty with "thing" categories, which are better characterized by overall shape than local appearance. In order to better capture the shape of "things," a few recent image parsing approaches [12, 16, 19] have attempted to incorporate sliding window detectors.Unfortunately, standard detectors based on HOG templates [2] or deformable part-based models (DPMs) [7] produce only a bounding box, whereas image parsing requires a pixel-level segmentation. Ladicky et al. [19] overcome this limitation by inferring a mask from a bounding box detection using GrabCut [29]. This automatic segmentation step can sometimes fail; moreover, it does not leverage the learned detection model, only the final bounding box. Guo and Hoiem [12] do not predict a mask from a bounding box but instead use the auto-context scheme of Tu and Bai [38] to directly incorporate the detector responses into their pixel-level parsing system.

To address the challenges of inferring object segmentations, we integrate region-based image parsing with the promising framework of *per-exemplar detectors* from Malisiewicz and Efros [24]. Per-exemplar detectors meet our need for pixel-level localization: when a per-exemplar detector fires on a test image, we can take the segmentation mask from the corresponding training exemplar and transfer it into the test image to form a segmentation hypothesis. By combining region-based parsing with per-exemplar detectors we are able to greatly increase the parsing accuracy of "thing" categories and achieve significant parsing accuracy improvements over our base parsing system on several challenging datasets.

SIFT Flow	Per-Pixel	Per-Class
Local labeling (Sec. 2)	74.1	30.2
MRF (Sec. 2)	76.2	29.1
Joint Geo/Semantic (Sec. 3)	77.0	30.1
Detector Combined (Sec. 4)	78.6	39.2
Liu et al. [21]	76.7	
Farabet et al. [5]	78.5	29.6
Farabet et al. [5] balanced	74.2	46.0
Eigen and Fergus [4]	77.1	32.5
Myeong et al. [25]	77.1	32.3

Table 1. SIFT Flow dataset: Semantic labeling accuracy and comparison to state of the art. “Per-Pixel” is the overall per-pixel classification rate and “Per-Class” is the average of the per-class rates.

2 Region-based nonparametric image parsing system

This section presents an overview of our region-based parsing system. It is based on a *lazy learning* philosophy, meaning that (almost) no training takes place offline; given a test image to be interpreted, our system dynamically selects the training exemplars that appear to be the most relevant and proceeds to transfer labels from them to the query. The following is a summary of the steps taken by the system for every query image.

1. Find a retrieval set of images similar to the query image.

2. Segment the query image into superpixels and compute feature vectors for each superpixel.

3. For each superpixel and each feature type, find the nearest-neighbor superpixels in the retrieval set according to that feature. Compute a likelihood score for each class based on the superpixel matches.

4. Use the computed likelihoods together with pairwise co-occurrence energies in an Markov Random Field (MRF) framework to compute a global labeling of the image.

Similarly to several other data-driven methods [21, 22, 30], our first step in parsing a query test image is to find a relatively small *retrieval set* of training images that will serve as the source of candidate superpixel-level matches. This is done not only for computational efficiency, but also to provide scene-level context for the subsequent superpixel matching step. A good retrieval set will contain images that have similar scene types, objects, and spatial layouts to the query image. In attempt to indirectly capture this kind of similarity, we use three types of global image features: spatial pyramid [20], gist [26], and color histogram. For each feature type, we add the n nearest neighbor training images to the retrieval set. Intuitively, taking the best scene matches from each of the global descriptors leads to better superpixel-based matches for region-based features that capture similar types of cues as the global features.

We wish to label the query image based on the content of the retrieval set, but assigning labels on a per-pixel basis as in [14, 21, 22] tends to be too inefficient.

LM+SUN	Per-Pixel	Per-Class
Local labeling (Sec. 2)	50.6	7.1
MRF (Sec. 2)	54.4	6.8
Joint Geo/Semantic (Sec. 3)	54.9	7.1
Detector Combined (Sec. 4)	61.4	15.2

Table 2. LM+SUN dataset: Semantic labeling accuracy. “Per-Pixel” is the overall per-pixel classification rate and “Per-Class” is the average of the per-class rates.

Instead, like [17, 23, 27], we choose to assign labels to superpixels, or regions produced by bottom-up segmentation. This not only reduces the complexity of the problem, but also gives better spatial support for aggregating features that could belong to a single object than, say, fixed-size square windows centered on every pixel in the image. For the second step, we obtain superpixels using the fast graph-based segmentation algorithm of Felzenszwalb and Huttenlocher [8] and describe their appearance using 20 different features similar to those of Malisiewcz and Efros [23], with some modifications and additions. Roughly speaking, these features describe size, shape, appearance, location, and texture of image regions.

Having segmented the test image and extracted the features of all its superpixels, we compute a log likelihood ratio score for each test superpixel and each class that is present in the retrieval set. We first compute the likelihood ratio score for each feature independently using a nonparametric density estimates of the features from the given class around the superpixel feature from the query image, and then combine scores from different features using the Naive Bayes assumption (see [33, 35] for details). At this point, we can obtain a labeling of the image by simply assigning to each superpixel the class with the maximum log likelihood ratio, which produces fairly competitive results (see “local labeling” in Tables 1 and 2).

Finally, we would like to enforce contextual constraints on the image labeling – for example, a labeling that assigns “water” to a superpixel completely surrounded by “sky” is not very plausible. Many state-of-the-art approaches encode such constraints with the help of conditional random field (CRF) models [9, 10, 14, 27]. However, CRFs tend to be very costly both in terms of learning and inference. In keeping with our nonparametric philosophy and emphasis on scalability, we restrict ourselves to contextual models that require minimal training and that can be solved efficiently. Therefore, we formulate the global image labeling problem as minimization of a standard MRF energy function defined over the field of superpixel labels $\mathbf{c} = \{c_i\}$:

$$J(\mathbf{c}) = \sum_{s_i \in SP} E_{\text{data}}(s_i, c_i) + \lambda \sum_{(s_i, s_j) \in A} E_{\text{smooth}}(c_i, c_j)\,, \tag{1}$$

where s_i is the ith superpixel region, c_i is the class assigned to that region, SP is the set of superpixels, A is the set of pairs of adjacent superpixels and λ is the smoothing constant. The dataterm ($E_{\text{data}}(s_i, c_i)$) is the likelihood ratio score weighted by the superpixel size. The smoothing term E_{smooth} penalizes adjacent superpixels that are assigned different labels, with a greater penalty to pairs of labels that co-occur less fre-

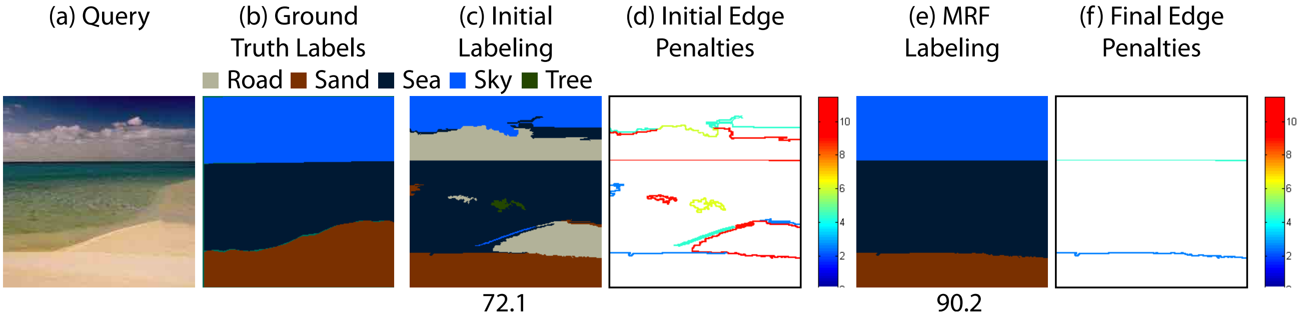

Figure 2. Our contextual edge penalty before and after we run our MRF optimization. Our contextual model successfully flags improbable boundaries between "sea" and "road".

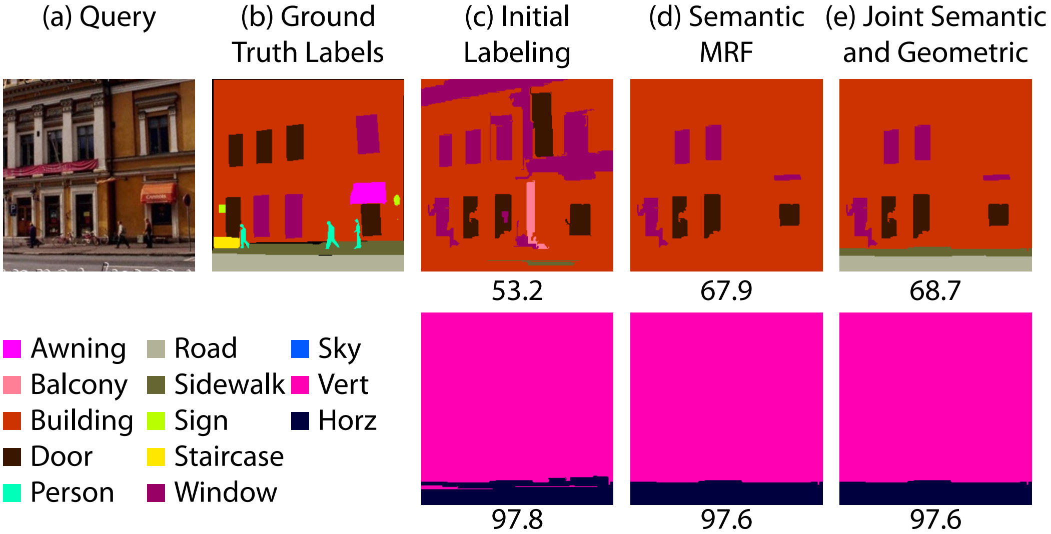

Figure 3. In the contextual MRF classification, the road gets replaced by "building," while "horizontal" is correctly classified. By jointly solving for the two kinds of labels, we manage to recover some of the "road" and "sidewalk" in the semantic labeling. Note also that in this example, our method correctly classifies some of the windows that are mislabeled as doors in the ground truth, and incorrectly but plausibly classifies the windows on the lower level as doors.

quently than others (see Figure 2 for an example). We perform MRF inference using the efficient graph cut optimization code of [1, 18].

We evaluate our system on two datasets. The first dataset, **SIFT Flow** [21], consists of outdoor scenes. It has 2,488 training images, 200 test images, and 33 labels. The second dataset, **LM+SUN** [35], was collected from the SUN dataset [39] and LabelMe [31]. It contains 45,676 images (21,182 indoor and 24,494 outdoor) and 232 labels. We use the split from our IJCV paper [35], which consists of 45,176 training and 500 test images. We measure two performance metrics: the per-pixel classification rate and the average per-class classification rate. The results for our system and other state-of-the-art systems are shown in Tables 1 and 2.

We initially presented this parsing system in ECCV 2010 [33]. In a subsequent IJCV paper [35], we explore the various components of this system in more detail and show how it can be used to parse video in a temporally consistent manner.

3 Simultaneous Classification of Multiple Label Types

To achieve more comprehensive image understanding and to explore a higher-level form of context, we consider the task of simultaneously labeling regions into two types of classes: semantic and geometric [10]. The notion of parsing an image into geometric classes was introduced by Hoiem et al. [17] and shown to be useful for a variety of tasks, such as rough 3D modeling and object location prediction. Like Hoiem et al. [17] and Gould et al. [10], we use three geometric labels – sky, horizontal, and vertical – although the sets of semantic labels in our datasets are much larger. In this work, we make the assumption that each semantic class is associated with a unique geometric class (e.g., "building" is "vertical," "river" is "horizontal," and so on) and specify this mapping by hand. We jointly solve for the fields of semantic labels (**c**) and geometric labels (**g**) by minimizing a cost function that is a simple extension of eq. (1):

$$H(\mathbf{c}, \mathbf{g}) = J(\mathbf{c}) + J(\mathbf{g}) + \mu \sum_{s_i \in SP} \varphi(c_i, g_i), \qquad (2)$$

where φ is the term that enforces coherence between the geometric and semantic labels. It is 0 when the semantic class c_i is of the geometric class type g_i and 1 otherwise. Figure 3 shows an example where joint inference over semantic and geometric labels improves the accuracy of the semantic labeling.

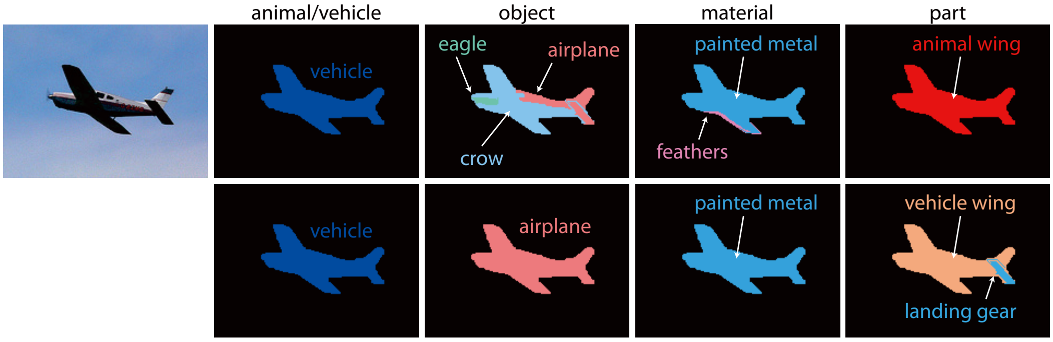

Figure 5. It's a bird, it's a plane! First row: single-level MRF inference for each label set in isolation. Second row: joint multi-level MRF inference. Animal/vehicle and material labelings are strong enough to correct the object and part labelings.

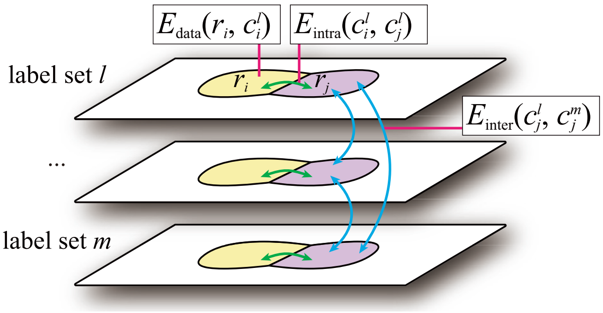
$E_{\text{data}}(r_i, c_i^l)$
$E_{\text{intra}}(c_i^l, c_j^l)$
label set l
r_i
r_j
$E_{\text{inter}}(c_j^l, c_j^m)$
...
label set m

Figure 4. Our multi-level MRF (eq. 3).

Label Set	Base	Single	Joint
Animal/veh.	92.8 (92.9)	92.8 (92.9)	92.8 (92.9)
Object	43.5 (41.8)	53.2 (50.5)	56.4 (36.7)
Material	51.8 (36.0)	54.1 (34.3)	53.9 (51.0)
Part	37.1 (11.2)	42.6 (11.7)	43.9 (12.3)

Table 3. CORE dataset results. "Base": the label with the maximum value of E_{data}. "Single": intra-label set smoothing only. "Joint": both intra- and inter-label set smoothing. The first number in each cell is the overall per-pixel classification rate, and the number in parentheses is the average of the per-class rates.

The results for jointly estimating the geometric and semantic labels are shown in Tables 1 and 2. The primary benefit of this system is a increase in the per-class performance, correcting the errors caused by the MRF smoothing. This work was presented in [33, 35].

Next, we generalize the two-level MRF objective function (2) to perform simultaneous inference over an arbitrary number of label sets. For example, the different label sets can correspond to object (semantic) labels, geometric labels, materials, parts, etc. We want to perform simultaneous inference over these label sets while enforcing constraints between labels from different sets for the same region: for example, "bird" (semantic) and "feathers" (material) labels are consistent, while "bird" and "metal" are not.

If we have n label sets, then we want to infer n labelings $\mathbf{c}^1, \ldots \mathbf{c}^n$, where $\mathbf{c}^l = \{c_i^l\}$ is the vector of labels from the lth set for every region $r_i \in R$. We can visualize the n labelings as being "stacked" together vertically as shown in Figure 4. "Intra-set" edges connect labels of neighboring regions in the same level just as in the single-level setup of Section 2, and "inter-set" edges connect labels of the same region from two different label sets. The MRF energy function on the resulting field is

$$\begin{aligned} E(\mathbf{c}^1, \ldots, \mathbf{c}^n) = & \sum_l \sum_{r_i \in R} E_{\text{data}}(r_i, c_i^l) \\ & + \lambda \sum_l \sum_{(r_i, r_j) \in A} E_{\text{intra}}(c_i^l, c_j^l) \\ & + \mu \sum_{l \neq m} \sum_{r_i \in R} E_{\text{inter}}(c_i^l, c_i^m), \end{aligned} \tag{3}$$

where $E_{\text{data}}(r_i, c_i^l)$ is the data term for region r_i and label c_i^l on the lth level, $E_{\text{intra}}(c_i^l, c_j^l)$ is the single-level smoothing term, $E_{\text{inter}}(c_i^l, c_i^m)$ is the term that enforces consistency between the labels of r_i drawn from the lth and mth label sets. Finally, the constants λ and μ control the amount of horizontal and vertical smoothing. E_{data} and E_{intra} are defined in the same way as in Section 2 with $E_{\text{intra}} = E_{\text{smooth}}$. As for the cross-level penalty E_{inter}, it is defined very similarly to to the intra-level penalty (E_{smooth}), based on cross-level co-occurrence statistics from the training set. If two labels often co-occur (e.g., "motorcycle" and "wheel") E_{inter} is low, while two label that rarely co-occur will have a high penalty.

To evaluate the above multi-level inference approach, we used the **Cross-Category Object Recognition (CORE)** dataset [6], which consists of 2,780 images and comes with ground-truth annotation for four label sets. The "objects" set has 28 different labels, of which 15 are animals and 13 are vehicles. The "animal/vehicle" label set designates each object accordingly. The "material" set consists of nine different materials and the "part" set consists of 66 different parts such as foot, wheel, wing, etc. The "material" and "part" sets have a many-to-many relationship with the object labels and both tend to be more sparsely labeled than the objects (i.e., not all of an object's pixels have part or material labels). Figure 5 shows one example of how our multi-level framework can correct for errors when each label set is parsed in isolation. Quantitative results are shown in Table 3. This work was presented in ICCV 2011 [34].

4 Parsing with Per-Exemplar Detectors

This section presents a hybrid image parsing system combining the region-based approach of Section 2 with sliding window object detectors. The overview of this system is given in Figure 6. First, we introduce our

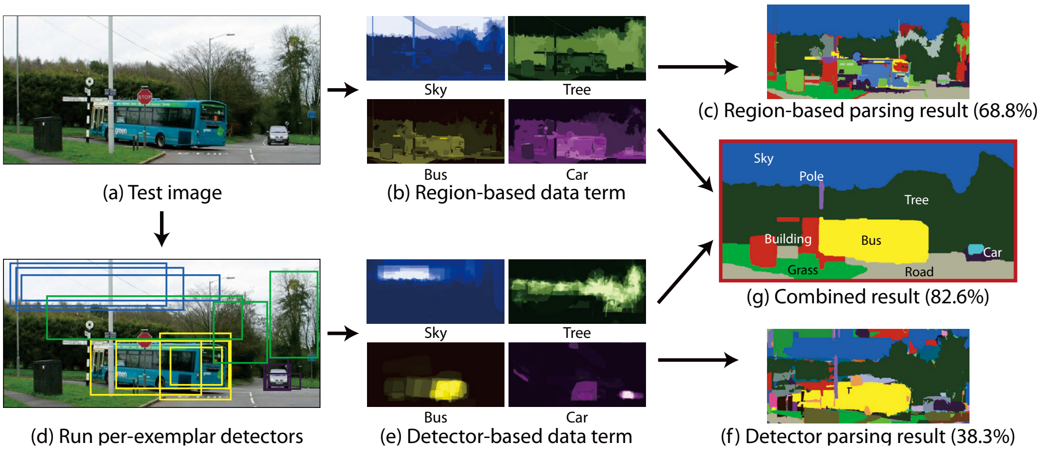
(a) Test image
Sky
Tree
Bus
Car
(b) Region-based data term
(c) Region-based parsing result (68.8%)
Sky
Pole
Tree
Building
Bus
Car
Grass
Road
(g) Combined result (82.6%)
(d) Run per-exemplar detectors
Sky
Tree
Bus
Car
(e) Detector-based data term
(f) Detector parsing result (38.3%)

Figure 6. Overview and sample result of our combined region- and detector-based approach. The test image (a) contains a bus – a relatively rare "thing" class. Our region-based parsing system computes class likelihoods (b) based on superpixel features, and it correctly identifies "stuff" regions like sky, road, and trees, but is not able to get the bus (c). To find "things" like bus and car, we run per-exemplar detectors [24] on the test image (d) and transfer masks corresponding to detected training exemplars (e). Since the detectors are not well suited for "stuff," the result of detector-based parsing (f) is poor. However, combining region-based and detection-based data terms (g) gives the highest accuracy of all and correctly labels most of the bus and part of the car.

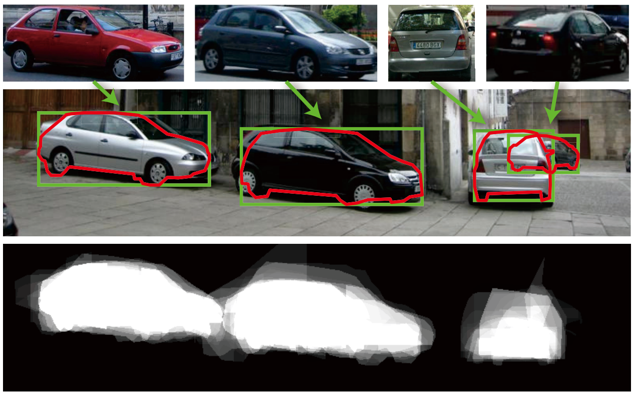

Figure 7. Computation of the detector-based data term. For each positive detection (green bounding box) in the test image (middle row) we transfer the mask (red polygon) from the associated exemplar (top) into the test image. The data term for "car" (bottom) is obtained by summing all the masks weighted by their detector responses.

new detector-based component and then we describe how we combine it with our region-based system.

Following the per-exemplar framework of Malisiewicz and Efros [24], we train a separate detector (HOG template) for each labeled object instance in our dataset. At test time, given an image that needs to be parsed, we first obtain a retrieval set of globally similar training images as in Section 2. Then we run the per-exemplar detectors associated with all ground-truth object instances in that retrieval set. For each positive detection we project the associated object mask into the detected bounding box (Figure 7). To compute the *detector-based data term* $E_{\mathrm{D}}(p, c)$ for a class c and pixel p, we simply take the sum of all detection masks from that class weighted by their detection scores. Figure 6(e) shows some detector-based data terms for the test image of Figure 6(a).

In addition, we define the *region-based data term* $E_{\mathrm{R}}(p, c)$ using the likelihood ratio score output by the system of Section 2. Now, for each pixel p and each class c in a test image, we end up with two data terms, $E_{\mathrm{R}}(p, c)$ and $E_{\mathrm{D}}(p, c)$. Next, we train a one-vs-all SVM for each class, each of which takes as input the both data terms and returns a final per-pixel scores for a given class c. Training data for each SVM is generated by running region- and detector-based parsing on the entire training set using a leave-one-out method: for each training image a retrieval set of similar training images is obtained, regions are matched to generate $E_{\mathrm{D}}(p, c)$, and the per-exemplar detectors from the retrieval set are run to generate $E_{\mathrm{D}}(p, c)$. To obtain the final labeling, we can simply take the highest-scoring label at each pixel, but this produces noisy results. We perform smoothing via a pixel-based MRF energy function similar to [21, 32].

The above system achieves state-of-the-art results on both the LM+SUN and SIFT-Flow datasets, as shown in Tables 1 and 2. Figures 8 and 9 show the per-class rates on the two datasets, while Figure 10 shows the output of various stages of our system on three test images. This work will appear in CVPR 2013 [36].

5 Discussion

Our parsing framework achieves fairly broad coverage on challenging large-scale datasets and it can leverage inter- and intra-label set context. However, there are a number of areas in which we are currently pursuing improvements. First, our detector based parsing framework is computationally expensive, especially during the training stage. Fortunately, training of per-exemplar detectors can be speeded up greatly using the whitened HOG method of Hariharan et al. [13]. Also, the per-exemplar detectors themselves are open-universe compatible (they can be trained individually and do not require re-training when new data is added), the SVM combination is not, as it requires batch training. We would like to train the combination incrementally in an online manner as new data arrives.

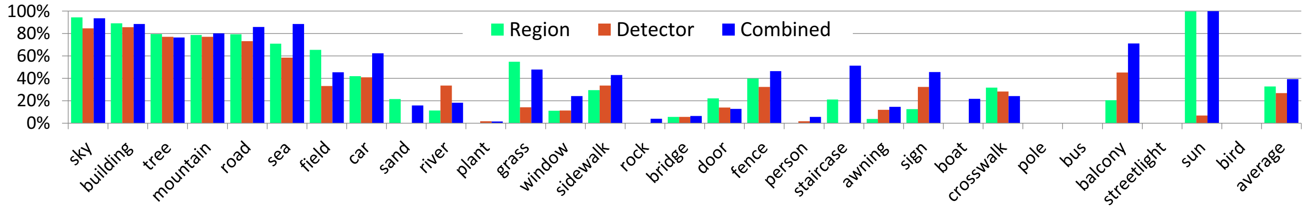

Figure 8. Classification rates of individual classes (ordered from most to least frequent) on the SIFT Flow dataset for region-based, detector-based, and combined parsing. All results include SVM and MRF smoothing.

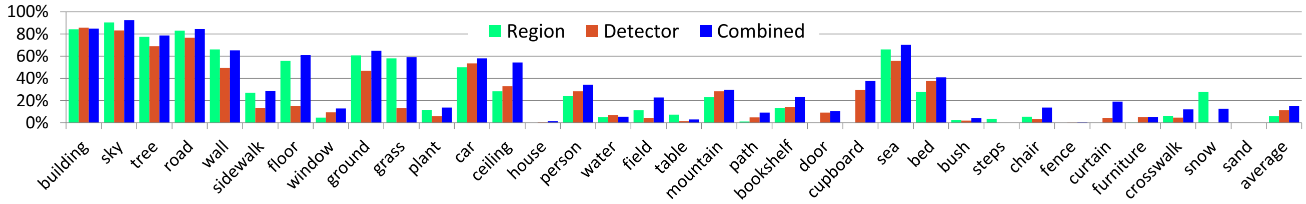

Figure 9. Classification rates of individual classes (ordered from most to least frequent) on the LM+SUN dataset for region-based, detector-based, and combined parsing. All results include SVM and MRF smoothing.

Furthermore, while our system achieves high pixel count accuracy, it currently does not identify object instances and does not produce accurate object boundaries. For example, nearby instances from the same class can get merged in the output, like the cars in Figure 10(b). We hope to leverage the per-exemplar framework to generate object instance hypotheses, which can be use during inference to improve object contours and separate different instances.

References

[1] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 26(9):1124–37, Sept. 2004.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[3] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, pages 1271–1278, June 2009.

[4] D. Eigen and R. Fergus. Nonparametric Image Parsing using Adaptive Neighbor Sets. In *CVPR*, 2012.

[5] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Scene Parsing with Multiscale Feature Learning, Purity Trees, and Optimal Covers. *Arxiv preprint arXiv:*, 2012.

[6] A. Farhadi, I. Endres, and D. Hoiem. Attribute-centric recognition for cross-category generalization. In *CVPR*, 2010.

[7] P. Felzenszwalb, D. Mcallester, and D. Ramanan. A Discriminatively Trained , Multiscale , Deformable Part Model. In *CVPR*, 2008.

[8] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient Graph-Based Image Segmentation. *IJCV*, 2(2):1–26, 2004.

[9] C. Galleguillos and S. Belongie. Context based object categorization: A critical survey. *CVIU*, 114(6):712–722, June 2010.

[10] S. Gould, R. Fulton, and D. Koller. Decomposing a Scene into Geometric and Semantically Consistent Regions. In *ICCV*, 2009.

[11] C. Gu, J. J. Lim, P. Arbel, and J. Malik. Recognition using Regions. In *CVPR*, 2009.

[12] R. Guo and D. Hoiem. Beyond the line of sight : labeling the underlying surfaces. In *ECCV*, 2012.

[13] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *ECCV*, 2012.

[14] X. He, R. S. Zemel, and M. A. Carreira-Perpinan. Multiscale Conditional Random Fields for Image Labeling. In *CVPR*, 2004.

[15] G. Heitz and D. Koller. Learning Spatial Context : Using Stuff to Find Things. In *ECCV*, pages 1–14, 2008.

[16] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*, pages 30–43, 2008.

[17] D. Hoiem, A. A. Efros, and M. Hebert. Recovering Surface Layout from an Image. *IJCV*, 75(1), 2007.

[18] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *PAMI*, 26(2):147–59, Feb. 2004.

[19] 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.

[20] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[21] C. Liu, J. Yuen, and A. Torralba. Nonparametric Scene Parsing via Label Transfer. *PAMI*, 33(12):2368–2382, June 2011.

[22] C. Liu, J. Yuen, and A. Torralba. SIFT flow: dense correspondence across scenes and its applications. *PAMI*, 33(5):978–94, May 2011.

[23] T. Malisiewicz and A. a. Efros. Recognition by association via learning per-exemplar distances. In *CVPR*, pages 1–8, June 2008.

[24] T. Malisiewicz and A. A. Efros. Ensemble of Exemplar-SVMs for Object Detection and Beyond. In *ICCV*, 2011.

[25] H. Myeong and K. M. Lee. Learning object relationships via graph-based context model. *CVPR*, June 2012.

[26] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 155:145–175, Jan. 2001.

[27] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in Context. In *ICCV*, pages 1–8, 2007.

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

[29] C. Rother, V. Kolmogorov, and A. Blake. GrabCut Interactive Foreground Extraction using Iterated Graph Cuts. *SigGraph*, 2004.

[30] B. C. Russell, A. Torralba, C. Liu, R. Fergus, and

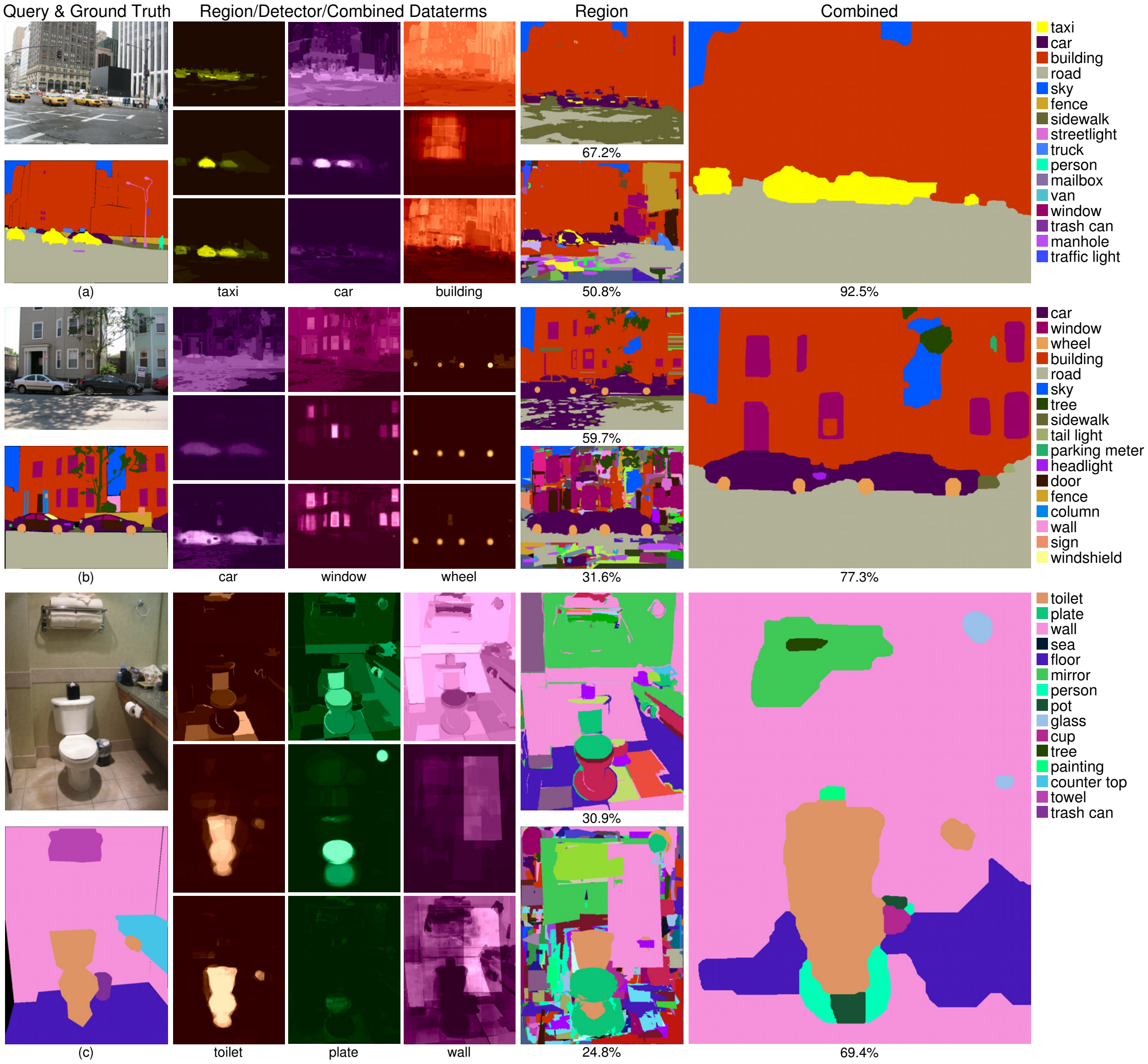

Figure 10. Example results on the LM+SUN dataset (best viewed in color). First column: query image (top) and ground truth (bottom). Second through fourth columns: region-based data term (top), detector-based data term (middle), and SVM combination (bottom) for three selected class labels. Fifth column: region-based parsing results (top) and detector-based parsing results (bottom) without SVM or MRF smoothing. Right-most column: smoothed combined output. The example in (a) has strong detector responses for both "car" and "taxi," and the SVM suppresses the former in favor of the latter. In (c), our system successfully finds the toilet. Note that both the region- and detector-based data terms assign very high likelihood of "plate" to the toilet bowl, but the SVM suppresses "plate" in favor of "toilet."

W. T. Freeman. Object Recognition by Scene Alignment. In *NIPS*, 2007.

[31] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe : a database and web-based tool for image annotation. *IJCV*, 77(1-3):157–173, 2008.

[32] 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.

[33] J. Tighe and S. Lazebnik. SuperParsing: Scalable Nonparametric Image Parsing with Superpixels. In *ECCV*, 2010.

[34] J. Tighe and S. Lazebnik. Understanding Scenes on Many Levels. In *ICCV*, pages 335–342, 2011.

[35] J. Tighe and S. Lazebnik. Superparsing Scalable Nonparametric Image Parsing with Superpixels. *IJCV*, 2012.

[36] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013.

[37] A. Torralba, R. Fergus, and W. T. Freeman. 80 Million Tiny Images: a Large Data Set for Nonparametric Object and Scene Recognition. *PAMI*, 30(11):1958–1970, Nov. 2008.

[38] Z. Tu and X. Bai. Auto-context and Its Application to High-level Vision Tasks and 3D Brain Image Segmentation. *PAMI*, pages 1–35, 2009.

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