Second-Order Constrained Parametric Proposals and Sequential Search-Based Structured Prediction for Semantic Segmentation in RGB-D Images

Dan Banica[1], Cristian Sminchisescu[2,1]
[1]Institute of Mathematics of the Romanian Academy, [2]Lund University
dan.banica@imar.ro, cristian.sminchisescu@math.lth.se

Abstract

We focus on the problem of semantic segmentation based on RGB-D data, with emphasis on analyzing cluttered indoor scenes containing many visual categories and instances. Our approach is based on a parametric figure-ground intensity and depth-constrained proposal process that generates spatial layout hypotheses at multiple locations and scales in the image followed by a sequential inference algorithm that produces a complete scene estimate. Our contributions can be summarized as follows: (1) a generalization of parametric max flow figure-ground proposal methodology to take advantage of intensity and depth information, in order to systematically and efficiently generate the breakpoints of an underlying spatial model in polynomial time, (2) new region description methods based on second-order pooling over multiple features constructed using both intensity and depth channels, (3) a principled search-based structured prediction inference and learning process that resolves conflicts in overlapping spatial partitions and selects regions sequentially towards complete scene estimates, and (4) extensive evaluation of the impact of depth, as well as the effectiveness of a large number of descriptors, both pre-designed and automatically obtained using deep learning, in a difficult RGB-D semantic segmentation problem with 92 classes. We report state of the art results in the challenging NYU Depth Dataset V2 [44], extended for the RMRC 2013 and RMRC 2014 Indoor Segmentation Challenges, where currently the proposed model ranks first. Moreover, we show that by combining second-order and deep learning features, over 15% relative accuracy improvements can be additionally achieved. In a scene classification benchmark, our methodology further improves the state of the art by 24%.

1. Introduction and Related Work

The problem of semantic segmentation in monocular images is of central importance in areas like robotics, human-computer interaction and scene understanding for large-scale indexing. For intensity images, significant progress has been achieved recently through work performed in association with the VOC Semantic Segmentation challenges[11], where high performing methods for boundary detection[36, 31], feature description and non-linear feature maps[4, 39, 37, 46, 33, 6], image segmentation [1, 5, 10, 32] as well as optimization and contextual reasoning [27, 38, 25, 15, 23, 7, 47] have been developed. Recently the use of deep feature extraction learning frameworks, trained on large-scale databases like Imagenet, has been shown to be effective not only for image classification [26], but also for semantic segmentation[14], where in conjunction with figure-ground proposal generation methods[5], impressive results have been achieved (the effectiveness of such regions descriptors will be analyzed in our proposed RGB-D framework, as well).

The scientific problem of three dimensional scene understanding from images, both quantitative[49, 12] and qualitative[21, 41], has a long standing research tradition in computer vision. Some of the more recent work has focused on the analysis of cluttered indoor scenes [22, 29, 16, 17, 20]. In this setup [29, 20] analyze the geometry of the rooms including surfaces and objects, whereas [17] reason about object functionality from the standpoint of a human user of the environment.

The existence of affordable and increasingly miniaturized time of flight and infra-red sensors like Kinect opens the possibility that RGB-D sensors will be embedded in any device, mobile or not in the near future. This creates scientific and technological opportunities for exploiting the RGB-D information for scene understanding and semantic segmentation, with potentially high gains in tasks that have been traditionally considered very challenging when performed based on intensity images alone. Range data has been extensively studied in the past, not only at the level of adapted descriptors like spin images[24] and 3D shape contexts[13] but also for shape modeling using, e.g., deformable superquadrics[30].

Besides the recent success for real-time human pose

estimation[42], Kinect has also spurred a wave of scene understanding research in robotics[39, 45] and computer vision[44, 39, 18, 34, 2, 3] with datasets[43, 28] recently made available. Our work relates to these recent RGB-D analysis approaches, and we will review them showing how we differentiate in methodology and focus. The NYU Depth Dataset V2 was introduced in [44], where the authors develop an expressive methodology for semantic segmentation by labeling merged superpixels while also inferring support relations between objects. Baseline approaches for semantic segmentation of RGB-D images were proposed in [43], where multiple alternatives were considered for the unary and pairwise terms inside a pixel-level CRF, with unary terms combining the output of a neural network applied on local descriptors and a depth-sensitive location prior; pairwise terms enforced smoothness while preserving depth discontinuities. In [39] a superpixel hierarchy is used, and the leaf superpixels are described using concatenated features (kernel descriptors) extracted from the entire path towards the root node of the segmentation tree. The work in [18] achieves excellent results for semantic segmentation after revisiting related problems such as boundary detection, bottom-up grouping and scene classification and extending the methodology to take advantage of depth information. The authors start with a hierarchy of non-overlapping (superpixel) partitions, and use the long-range amodal completion of surfaces for better region grouping.

Our methodology differentiates from the above approaches in our multiple figure-ground proposal generation based on parametric max-flow extended to use intensity and depth information[1], as well as in the feature description, second order pooling, and inference procedure used, which is adapted to handle RGB-D models with many categories and scenes where many instances are present, at widely varying spatial scales. Our pooling process operates over descriptors that capture both appearance (e.g. SIFT [35]) and geometry (e.g. spin images [24]). Besides the pooled local descriptors we also extracted point cloud features to coarsely characterize the aspect and size of each region. Also, we extracted features from a large convolutional neural network trained for image classification on ImageNet. In the experiments we report the performance of the above features individually and also show the benefits of using them together to jointly describe each region. A class label is assigned to each segment by learning linear category models, one per class, trained to predict the overlap (IoU) between the segment and the best matching object of that class. Finally, a principled search-based structured prediction inference procedure is defined in order to resolve conflicts between overlapping segments which were assigned different labels, and to generate a final per-pixel segmentation. We differentiate from previous sequential search-based scene parsing procedures (*e.g.* [40, 48]) in our training procedure derived from SEARN [8], which allows optimizing decisions under the true metric, using partial segmentations as intermediary states and resolving conflicts between the overlapping regions, as opposed to *e.g.* inferring labels of non-overlapping superpixels. We analyze the effectiveness of integrating depth, as well as the proposed solutions at each stage of this pipeline, perform analysis of alternative features including those obtained from deep learning, and show that in the challenging NYU Depth Dataset V2 [44], extended for RMRC 2013 and RMRC 2014 Indoor Segmentation Challenges, the proposed model ranks first.

The rest of the paper is organized as follows: §2 presents how depth data is used in order to improve the generation of figure-ground segmentations within parametric max-flow models, §3 illustrates the procedure used for assigning a label to each segment, while §4 describes the sequential search-based structured prediction inference and learning procedure we propose in order to resolve conflicts between overlapping segments and obtain final per-pixel labels for the entire image. Experiments follow in §5. We conclude in §6.

2. Parametric Generation of Figure-Ground Proposals

In contrast to methodologies that compute hierarchical, non-overlapping partitions of the image into multiple regions, our approach relies on generating multiple overlapping figure-ground segmentations, systematically, based on parametric max-flow solvers. We focus on constrained parametric min cuts models CPMC[5] generalized to take advantage of intensity and depth information (CPMC-RGBD). We rely on simple spatial energy models based on attention mechanisms that allow us to solve for all breakpoints (segmentation solutions), corresponding to different locations and spatial scales, in polynomial time. The idea is to 'fixate' at different spatial locations, set up constraints such that a fixated location is assigned to the foreground, and elements on the boundary of the image are assigned to the background, then solve for the set of binary partitions that can be obtained under such constraints. Because solutions obtained at different fixation points may overlap, or may have low quality, skewed shape statistics, a ranking process ensures that only a valid and compact subset is retained. The ranker (in our case a linear regressor) is trained to distinguish between those segments that exhibit the regularities of real-world objects (e.g. continuity, convexity, Euler structure, etc.) and the ones that do not. This 'objectness' criteria is category independent: the ranker is

[1] Note that in parallel with the initial versions of our work, ideas based on our earlier RGB-based constrained parametric min cuts (CPMC) [5] and second-order pooling (O2P) [4] have also been used for RGB-D data in [34]. In any case, notice however, that [34] address the different task of 3D object detection, providing methodology to assign labels to 3D cuboids, instead of a pixel-level segmentation, as our focus in this work.

trained using a large variety of shapes belonging to many visual categories. Following duplicate elimination and hypothesis scoring, a Maximal Marginal diversification stage ensures that the pool of solutions obtained contains good quality configurations that are sufficiently different from each other.

The figure-ground segmentation proposals are generated by solving a family of optimization problems for spatial energies of the form:

$$E^{\lambda}(L) = \sum_{x} D_{\lambda}(l_x) + \sum_{x,y \in \mathcal{N}(x)} V_{xy}(l_x, l_y) \qquad (1)$$

where L is a labeling of the pixels in the image into foreground or background, $\mathcal{N}(x)$ is the neighborhood of a particular pixel/node x, $\lambda \in \mathbb{R}$ selects the problem instance to be solved, the unary term D_{λ} defines the cost of assigning a particular pixel to the foreground or the background, and the pairwise term V_{xy} penalizes the assignment of different labels to similar neighboring pixels.

In order to incorporate depth data, our CPMC-RGBD method modulates the pairwise term of the spatial model to account for both intensity and depth discontinuities, resulting in a more accurate pool of segments (see fig. 1 for qualitative results). The intensity-based pairwise term V_{xy} in eq. 1 has the following form: $V_{xy}(l_x, l_y) = \exp\left[-\frac{\max(B_{\mathcal{I}}(x), B_{\mathcal{I}}(y))}{\sigma^2}\right]$ when two neighboring pixels x, y are assigned different labels, where $B_{\mathcal{I}}$ is the output of a generalized, trained contour detector [31, 36] computed for the image $\mathcal{I}$ at a given pixel. In order to fuse depth information, we relied on two boundary probability maps, one based on the RGB information in the image and the other using the depth image. We define an augmented penalty which has the following form when neighboring pixels x, y are assigned different labels:

$$V_{xy}(l_x, l_y) = \exp\left[-\frac{\max(B_{\mathcal{I}}(x), B_{\mathcal{I}}(y), B_{\mathcal{D}}(x), B_{\mathcal{D}}(y))}{\sigma^2}\right] \qquad (2)$$

where $B_{\mathcal{D}}$ is the output of a global contour detector [31, 36] on the depth image. The effects of the proposed boundary fusion scheme are illustrated in fig. 2 where it can be seen that we can adaptively select useful boundaries using both RGB and depth cues.

By solving for $\min_{\lambda,L} E^{\lambda}(L)$ of the sub-modular energy using parametric max-flow, we systematically obtain an entire family of nested solutions in polynomial time (the nesting property of the solutions for this model enables an efficient solver for all breakpoints). For imaging models, the nesting property also ensures that solutions are obtained at different spatial scales in the image – provided that our 'attention mechanism' operates over a sufficiently fine grid, both small and large objects are usually covered quite well. The segments are ranked using a class independent predictor, based on the object-like regularities that each region exposes. We use this category-independent ranker to retain only the top $K = 500$ scoring hypotheses for further processing.

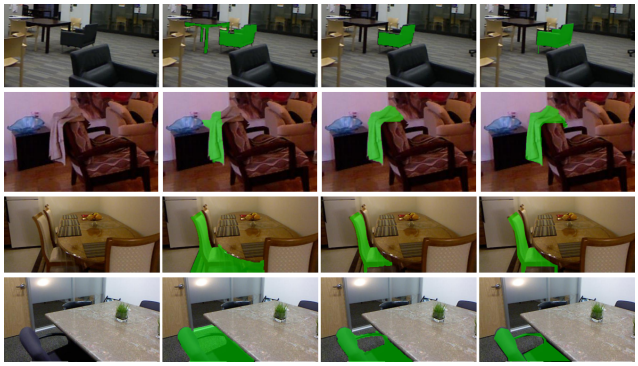

Figure 1. By combining depth and intensity cues we can significantly improve the quality of the figure-ground proposal. Left to right: original image, best segment obtained from constrained parametric max-flow on intensity images (CPMC), best segment from CPMC-RGBD that combines intensity and depth information, and ground truth. The images are from the NYU Depth Dataset V2 [44].

Fig. 1 illustrates how better segment pools are obtained by fusing RGB and depth information in CPMC-RGBD. Notice that thin structures (considering the detail available at that spatial scale) and fine details of objects are captured extremely well – see for instance the legs or the arm rest of chairs. This is promising for robotic RGB-D sensing systems that would be capable to both recognize and manipulate objects in the long run. Quantitatively, the improvement due to the usage of depth is also significant (§5).

3. Description and Recognition of Regions

3.1. Second-Order Pooling Over Local RGB-D Descriptors

To characterize a proposal region, we use local descriptors that capture both the appearance and the depth information available in the RGB-D images. Local descriptors extracted inside the region are aggregated using Second Order Pooling (O2P) [4]. O2P introduces multiplicative second-order analogues of average pooling that together with additional non-linearities (matrix logarithm, power normalization) produce good predictors without the need of going through a feature coding step.

We pool local features characterized by say, M descriptors, $X = (\mathbf{x}_1, \ldots, \mathbf{x}_M)$, $\mathbf{x} \in \mathbb{R}^n$, extracted over patches centered at image locations sampled inside the region proposal R, to form global descriptors based on second-order statistics. We will exploit multiplicative second-order interactions (e.g. outer products), with average operators. We define *second-order average-pooling* (2AvgP) as the matrix:

Figure 2. Depth and intensity boundary complement each other. Left to right: original image, boundaries extracted from the intensity image, boundaries extracted from the depth image, boundaries resulting from the fusion of RGB and depth information, *cf* (2) with winning channel shown.

$$\mathbf{G}_{avg}(R) = \frac{1}{M} \sum_i \mathbf{x}_i \cdot \mathbf{x}_i^\top, \tag{3}$$

As the second order pooling operator constructs a symmetric positive definite matrix, we will use the log-Euclidean metric adapted for this space. We apply this operator on the second-order statistics $\mathbf{G}_{avg}$ of each region proposal R_j, generated using CPMC-RGBD:

$$\mathbf{G}^{log}_{avg}(R) \leftarrow \log\left(\mathbf{G}_{avg}(R)\right), \tag{4}$$

The logarithm is obtained using the Schur-Parlett algorithm which takes $\mathcal{O}(n^3 \div n^4)$ operations depending on the distribution of eigenvalues of the input matrices.

Our pooling process considers both RGB and depth information. We first pool features that have proven effective for RGB data [4] – SIFT, masked SIFT, Local Binary Patterns (the LBP descriptor). In order to exploit the additional depth information available, we pool over spin images[24], masked spin images and SIFT, masked SIFT, and Local Binary Patterns applied to the depth image. The main differences between the masked and non-masked version of a descriptor occur at those points near the boundaries of the region, where the spatial support of the local descriptor may include fragments outside the current region, belonging to other objects – choosing to ignore the points outside the current region leads to the masked version of the descriptor. The 3D local descriptors are further enriched using location and color information.

3.2. Structural 3D Point Cloud Features

In order to better characterize the structure of a region proposal, we additionally extract a series of measurements from the 3D bounding box of the point cloud associated to it. We characterize the 3D bounding box of the region proposal by 11 numbers: volume, surface, diagonal, perimeter (sum of all side lengths), min side length, median side length, max side length, the length of each side along the 3 axes, and aspect ratio (min side / max side). Fitting a bounding parallelepiped to a region point cloud exactly may not produce desirable results due to noise. Therefore in order to achieve robustness we ignore a fixed percent extremal points along each of the 3 axes. This outlier percent was varied (0%, 2.5%, 5%, 7.5%), to generate the 11-dimensional feature vector for each threshold process. We combined the 4 levels to obtain a 44 dimensional descriptor for the point cloud, then let the classifier decide what represents a good threshold.

3.3. Confidence Models for Region Categories

The second order RGB-D descriptors and the 3D point cloud features described in the previous two sections are concatenated and used as a joint region descriptor. For each category we train linear regression models to predict the overlap between a region and the best-matching objects of each class – one predictive model is trained for each category. The data used for building the predictive category models is composed of the features extracted on the ground truth masks from the training set along with the K masks generated by CPMC-RGBD for each training image, with their true Intersection over Union (IoU) overlap with the ground truth. For the ground truth masks the target value will be 1 for the predictive model associated to the specific class of the object, and 0 for all other models, whereas for the imperfect CPMC-RGBD segments the target output will be a value in the $[0, 1]$ interval.

At test time, we assign a class label to each of the K retained masks by running all category predictors and choosing the class with maximal estimated overlap. The regression model naturally provides a useful confidence measure, for each proposal and visual category. While this provides a

Figure 3. Sequential Search-Based Structured Prediction procedure at test-time. We iteratively apply the learned policy and select the action predicted as most suitable given current state.

decision at the level of regions considered in isolation, such regions may overlap. In order to construct the final solution, the predicted labels of regions together with their confidence will be used within a sequential inference process that resolves conflicts and assigns labels for entire image.

4. Sequential Search-Based Inference

At this point, for a given test image, we have K overlapping object-level proposals which have been independently labeled to visual categories using the methodology just described. We also have confidences for estimates. Our objective is to generate a per-pixel labeling. This is not straightforward because the K object-level proposals overlap.

We formulate the solution in terms of a principled sequential framework which involves learning a policy to optimally select among a set of actions until a final state is reached. The quality of a decision can be evaluated either immediately, or by considering the long term effect of the action, after multiple decisions have been made. The procedure described in this section shares connections with reinforcement learning and is in the spirit of Search-based Structured Prediction (SEARN) [8]. We are not aware of an application of such principles for the task of semantic segmentation. Formally, we learn a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$, where $\mathcal{S}$ represents the set of states (partial semantic segmentations of an image in our case) and $\mathcal{A}$ represents the set of actions.

In practice our method generates a complete image segmentation by selecting one labeled region at a time. The partial semantic segmentations reached during this process represent the intermediary states (see fig. 3). In our implementation the actions are represented by candidate labeled regions to be added next, and there is also a special 'stop action' which determines finalization of the process.

The generic training method is described in Algorithm 1 (see [8]). We pursue two instantiations of this framework. In both situations, a state is a partially segmented image, represented by a set of regions selected so far, along with their associated classes: $s = \{(r_i, c_i) | i = \overline{1..k}\}$. We first generically describe the common elements of the two instantiations and will afterwards detail the specific elements of each.

In order to train a policy to optimally select an action available in a given state (partial segmentation), we gather multiple cost-sensitive training examples. Ideally, these

Algorithm 1 Learn a policy for sequential inference

Require: initial policy π_i, train set $\mathcal{D}$
Ensure: π – the learned policy
Initialize current policy $\pi \leftarrow \pi_i$
while termination criterion not reached **do**
  Initialize the set of cost-sensitive examples $\mathcal{E} \leftarrow \emptyset$
  for each $d \in \mathcal{D}$ **do**
    Apply π, generate a seq. of states $(s_0, s_1, ...s_n)$
    for each partial segmentation s_i **do**
      for each action α_j available in s_i **do**
        Compute features $\theta = \theta(s_i, \alpha_j)$
        Compute loss $l = l(\alpha_j, s_i, \pi)$ for action α_j
        Add (θ, l) to $\mathcal{E}$
      end for
    end for
  end for
  $\pi' \leftarrow$ new policy trained using the examples in $\mathcal{E}$
  Current policy $\pi \leftarrow$ interpolation of π and π'
end while

training examples should be generated from intermediary partial segmentations which are similar in nature to those which will be encountered at test time. However, it is not straightforward to achieve this, since at test time the intermediary states will be generated by applying the learned policy. To this end, an iterative procedure is used for policy training. We start from an initial policy π_i (which can be defined using ground-truth), generate training examples from partial segmentations obtained by applying this policy, train a new policy, and repeat using this new policy.

The features $\theta(s_i, \alpha_j)$ derive information from both the current partial segmentation s_i and also from the current candidate action α_j – their aim is to capture how suitable it is to add a candidate labeled segment given current partial segmentation. In our implementation we included co-occurrences between current candidate label and previously selected labels, number of segments, number of distinct labels selected and confidence for the candidate segment.

The loss $l(\alpha_j, s_i, \pi)$ models the effect of performing candidate action α_j in the state s_i. One benefit of the framework is that it allows evaluating the long-term effect of the selected action (by analyzing the final segmentation which results after applying the current policy) under the metric that we optimize (*e.g.* mean per class Jaccard index).

We detail two instances of this generic procedure.
Decide-region. Here, the actions available in a given state s are $\mathcal{A}_s = \{(r_j, c_j) | j = \overline{1..l}\} \cup \{\alpha_{stop}\}$, where each (r_j, c_j) pair represents a region with an associated class – these are the segments which do not overlap[2] with the segments already selected in current partial segmentation s. Selecting such an action advances the system in a new state, which

[2]We permit minor overlaps – we only reject candidates whose IoU with previously selected regions is above a fixed 0.05 value

is obtained by adding the new (r_j, c_j) labeled region to the set of already selected labeled regions. There is also a special action, denoted α_{stop} for which the final state is reached when selected – in this case, the current labeled regions represent the segmentation of the input image. The initial policy π_i is based on ground-truth (*i.e.* always choose the best available segment) and a fixed number of iterations (10) is used as termination criterion.

Decide-continuation. In this instantiation the actions available in a given state s are $\mathcal{A}_s = \{\alpha_{continue}, \alpha_{stop}\}$. As before, when α_{stop} is selected, the final state is reached. When $\alpha_{continue}$ is selected, a new state is generated by adding the region with the highest class specific estimated confidence (§3). During training, in each intermediary state we generate a training example. The loss for stop action is positive (*i.e.* continuation is encouraged) if the best possible state (evaluated under the metric to be optimized) can be reached later by sequentially applying the $\alpha_{continue}$ action.

5. Experiments

Our experiments were conducted on the NYU Depth Dataset V2 [44], which contains 1449 RGB-D images. We model 92 object classes for semantic labeling, each being found at least 50 times in the NYU Depth Dataset V2.

We also show results on two extensions of this dataset, introduced for the RMRC 2013 and RMRC 2014 Indoor Segmentation Challenges, held during ICCV 2013, respectively ECCV 2014, where our method currently ranks first.

In our implementation the spin images pooled using O2P were represented by 16×16 2D histograms, extracted at two spatial scales – considering points within a radius of 0.3 respectively 0.5 meters. The RGB-based local descriptors (SIFT and LBP) were computed using the same parameters as in the publicly available implementation of O2P. In the PCA reduction step we retained 2,500 dimensions from the pooled spin images and 2,500 dimensions from the pooled masked spin images, along with the 12,500 dimensions retained from the descriptors which use RGB information (SIFT and LBP). When pooling SIFT, masked SIFTs and LBP descriptors on the depth image we used the same parameters as for RGB, but when reducing the dimensionality of the descriptors using PCA we retained 2,500 dimensions from each descriptor type (instead of retaining 5,000 dimensions as in RGB for each variant of the SIFT descriptors – masked/not masked).

We have also experimented with 'deep features' extracted from a large convolutional neural network trained for image classification on ImageNet. We followed the procedure and implementation from [14], without the fine-tunning step, using the network architecture defined in [26] which resulted in a 4096 dimensional feature vector.

We next analyze the effects of various components of the system, at each stage. Unless otherwise indicated, the results reported below are obtained on the test set of NYU Depth V2, using the standard train-test split which consists of 795 training images and 654 testing images.

CPMC	CPMC-RGBD	CPMC+[9]	Upp-bnd.
55.6 (707)	59.1 (1166)	57.7 (762)	68.1

Table 1. Integrating RGB and depth cues generates improved figure-ground segmentations. The values represent the average IoU measures over ground-truth objects for the best-matching proposal using different methods. The numbers in the parantheses specify the average number of segments generated per image. We present the scores for the CPMC algorithm ([5]), the CPMC-RGBD algorithm (§2), the score obtained by integrating depth into CPMC-RGBD using an edge detector trained on RGB and Depth simultaneously ([9]), upper bound generated assuming that perfect (gt) boundaries are available.

Parametric Generation of Figure-Ground Proposals: We have generated proposals using a regular 'attention model' based on a 5x5 grid of seeds, and constraints placed as described in §2. We first investigated the impact of depth in the generation of the segment pool. We show qualitative results in fig. 1 and quantitative ones in table 1.

Description and Recognition of Regions: After extracting multiple figure-ground segment proposals based on RGB-D, each of them is categorized, with confidence, using the procedure described in §3. We retained $K = 500$ segments from each testing image (the highest-scoring regions according to a category-independent ranker). For training we used both the clean ground truth masks and noisier automatically generated segment proposals. We observed only marginal improvements when training with more than 300 masks per image – therefore we only retained 300 segments for training, which are passed to category-specific predictors, along with ground truth segments. Notice that we use 300 proposals in training and 500 in testing. There is no inconsistency as these numbers need not be the same – in practice we have also experimented with mixed regular and irregular grids where we made sure that we always placed seeds on ground truth objects in training, but this strategy did not produce significantly better pools than the ones based on a regular 5x5 seeding grid.

We will extensively analyze the performance of the segment descriptors constructed based on both RGB and depth information. We report intermediary results as well since the inference process that estimates per-pixel segmentations involves steps which are in turn prone to error.

Labeling Ground Truth Segments: We begin by analyzing the performance of our descriptors on the clean ground truth segments from the NYU Depth V2 test set. Results are shown in table 2. Interestingly, the pooled depth descriptors performed better than the RGB descriptors. However, their combination significantly boosted the score, confirming that indeed complementary information is present in the depth and intensity channels, and our model can leverage it.

Deep features	O2P on local descriptors						PCF	O2P + PCF	O2P + deep feats.
	all RGB	Depth features				all RGB-D			
		spin imgs	SIFT depth	LBP depth	all depth				
45.43	55.98	47.04	52.39	40.84	57.22	62.95	16.46	62.94	64.54

Table 2. Accuracy of different RGB and depth descriptors in labeling the ground-truth segments on the NYU Depth V2 test set.

Labeling Figure-Ground RGB-D Segment Proposals: We next analyzed the behavior of the descriptors considering the segments generated by our parametric solver operating on RGB-D channels. This aims to analyze robustness of descriptors with respect to imperfections in segmentation. Categorizing segments individually is the final step before proceeding to inference described in §4 where overlapping segments compete for pixel labeling. The performance of labeling imperfect segments is shown in table 3.

Semantic Segmentation: In table 5, we report the end-to-end performance using various descriptors for labeling segments. The metric is the one used in the RMRC Indoor Semantic Segmentation Challenge held during ICCV 2013 – mean recall per class.

RMRC 2013 results: In table 6 we show the scores of our segmentations, which were uploaded on March 06, 2014 on the RMRC test server. These segmentations were generated using only the pooled local descriptors (O2P).

RMRC 2014 results: Table 7 shows the performance of the winning entries for the recent RMRC 2014 Indoor Segmentation Challenge. This competition used a different evaluation metric – mean intersection over union scores per class. The set of semantic labels consists of 23 frequently occurring object classes. Here, motivated by the effectiveness of our categorization with confidence methods in §3, we retained full segment pools for processing, bypassing the category-independent ranker.

Evaluation of the inference procedure: In §4 we described a generic sequential framework for selecting a subset of non-overlapping segments. In table 4 we compare two instantiations of the generic framework (§4) with a baseline that consists of running the Decide-region with a single iteration – *i.e.* for the baseline we train with partial segmentations that consist of subsets of best available segments (ground truth based initial policy). The method with highest score (denoted as Decide-continuation) was evaluated on RMRC 2014 test server and achieved a Jaccard index score of 0.32 (table 7).

Failure cases: Errors occur at different stages of our pipeline, but quantitatively, larger gains can be achieved by further improving class predictions of segments (§3), as opposed to *e.g.* having perfect candidate regions and applying the current labeling methods on top. In general small objects are problematic both for segment proposal and for classification stage; thin planar background elements are often confused with pictures (*e.g.* fig. 4, row 2); textureless elements are difficult in general (*e.g.* fig. 4, row 3 - light projected on wall classified as window).

Method	Score
Decide-Region, single iteration	34.65
Decide-Region	37.33
Decide-Continuation	40.39

Table 4. Segmentation scores achieved with different inference procedures. The results are obtained on the test set of NYU Depth V2 dataset consisting of 654 images. The metric used is the mean Jaccard index (intersection over union) per class. The label set is represented by the 23 object classes which were used in the RMRC 2014 Indoor Segmentation Challenge.

Method	Classes won	Average score
Gupta et al. [18]	32	23.98
Silberman et al. [44]	29	21.31
Ren et al. [39]	22	17.52
Ours	**39**	**24.61**

Table 6. Semantic segmentation performance under the average recall per class metric, for 92 classes. The reported results are obtained on the RMRC 2013 test set (an extension of the NYU Depth V2 dataset) after uploading our results on the evaluation server. The metric is the average recall per class ('average score' column). We also report the number of classes where each method achieves the highest score (in case of ties, one point is added for each method achieving the highest score). The uploaded method uses the O2P descriptors (without deep learning features)

Method	Score
Ours	**0.32**
Gupta et al. [19]	0.30

Table 7. Winners of the RMRC 2014 semantic segmentation challenge. Each pixel in each image is labeled with one of the following 23 classes: background, bathtub, bed, blinds, cabinet, ceiling, chair, counter, curtain, desk, dresser, floor, night stand, picture, pillow, refrigerator, shelves, sofa, table, television, toilet, wall, window. The metric is the Jaccard index: the mean of the per-class intersection over union scores.

Scene classification: Motivated by the accuracy of the pooled local descriptors we also tackled the problem of scene classification (also studied in [18]) and investigated the improvements that resulted by adding depth information. We applied the second-order pooling machinery on top of the same local descriptors presented in §3.1, that capture both appearance and depth. The pooling of local descriptors was done in a spatial pyramid, homogeneously (no segmentation proposals) by dividing the entire image in 1, 2×2,

Deep features	O2P on local descriptors						PCF	O2P + PCF	O2P + deep feats.
	all RGB	Depth features				all RGB-D			
		spin imgs	SIFT depth	LBP depth	all depth				
61.69	56.87	46.01	54.90	46.27	59.35	65.22	12.87	65.54	67.15

Table 3. Accuracy for labeling figure-ground RGB-D proposals extracted automatically, on the NYU Depth V2 test set. The correct label of a proposal is assumed to be the label of the ground truth object that mostly overlaps that segment. Only segments that have at least 50% overlap with a ground truth object are considered.

Deep features	O2P on local descriptors						PCF	O2P + PCF	O2P + deep feats.
	all RGB	Depth features				all RGB-D			
		spin imgs	SIFT depth	LBP depth	all depth				
20.80	18.68	13.13	16.64	11.06	20.49	24.68	3.28	24.10	29.03

Table 5. Semantic segmentation performance on the NYU Depth V2 test set under the average recall per class metric.

Figure 4. Sample semantic segmentations generated by our system. Left to right: RGB image, ground truth semantic segmentation, our segmentation.

4×4 grids. State of the art results were achieved, results are shown in table 8.

Class	[18]	RGB	Depth	RGB-D
bedroom	79	79.06	78.01	82.72
kitchen	74	65.09	60.38	75.47
living room	47	73.83	33.64	75.70
bathroom	67	89.66	81.03	96.55
dining room	47	96.36	50.91	96.36
office	24	63.16	13.16	71.05
home office	8.3	70.83	0.00	62.50
classroom	48	69.57	52.17	82.61
bookstore	64	100.00	72.73	100.00
others	15	85.37	39.02	95.12
mean diag. cm.	47	79.29	48.11	**83.81**
avg. accuracy	58	77.52	55.81	**82.42**

Table 8. Scene classification performance on the NYU Depth V2 test set, measured using the mean-diagonal of the normalized confusion matrix (average precision per class) and average classification accuracy. The 'RGB' column shows results obtained using descriptors that use RGB data only (SIFT, LBP), pooled using O2P; the 'Depth' column gives results using only pooled local descriptors, SIFT, LBP, spin images computed on depth channels.

6. Conclusions

We have presented a semantic segmentation methodology for RGB-D data, where we have focused on cluttered indoor scenes containing many visual categories. Our approach is based on a parametric figure-ground intensity and depth-constrained proposal process that systematically generates spatial layout hypotheses at multiple locations and scales in the image followed by a novel, optimal sequential inference algorithm that integrates conflicting proposals into a complete scene estimate. We contribute by: (1) generalizing parametric max flow figure-ground methodologies to take advantage of intensity and depth information, (2) region description methods based on second-oder pooling over multiple features constructed using both intensity and depth channels, (3) a principled search based structured prediction inference and learning process that can select regions sequentially towards complete scene estimates, (4) evaluation of the impact of depth, as well as the effectiveness of a large number of descriptors, both pre-designed and automatically obtained using deep learning, in a difficult RGB-D semantic segmentation problem with 92 classes. We report state of the art results in the challenging NYU Depth Dataset V2 [44], extended for the RMRC 2013 and RMRC 2014 Indoor Segmentation Challenges, where the proposed model ranks first. By combining second-order and deep learning features, accuracy improvements in excess of an additional 15% can be attained. In a RGB-D scene classification benchmark, our methodology further improves the state of the art by 24%.

Acknowledgements

This work was supported in part by CNCS-UEFISCDI under CT-ERC-2012-1 and PCE-2011-3-0438.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *PAMI*, 33(5), 2011. 1

[2] D. Banica, A. Agape, A. Ion, and C. Sminchisescu. Video object segmentation by salient segment chain composition. In *Inference for probabilistic graphical models (PGMs) Workshop, ICCV*, 2013. 2

[3] D. Banica and C. Sminchisescu. Cpmc-3d-o2p: Semantic segmentation of rgb-d images using cpmc and second order pooling. *Oral presentation at the Reconstruction Meets Recognition Challenge (RMRC) Workshop, ICCV, December, arXiv preprint:1312.7715v1*, 2013. 2

[4] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*. 2012. 1, 2, 3, 4

[5] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *PAMI*, 34(7), 2012. 1, 2, 6

[6] R. G. Cinbis, J. Verbeek, and C. Schmid. Segmentation Driven Object Detection with Fisher Vectors. In *ICCV*, 2013. 1

[7] C. Dann, P. Gehler, S. Roth, and S. Nowozin. Pottics–the potts topic model for semantic image segmentation. In *Pattern Recognition*. 2012. 1

[8] H. Daumé III, J. Langford, and D. Marcu. Search-based structured prediction. *Machine learning*, 75(3), 2009. 2, 5

[9] P. Dollár and C. L. Zitnick. Structured forests for fast edge detection. In *ICCV*, 2013. 6

[10] I. Endres and A. Hoiem. Category independent object proposals. In *ECCV*, 2010. 1

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 1

[12] D. Forsyth, J. Mundy, A. Zisserman, C. Coelho, A. Heller, and C. Rothwell. Invariant descriptors for 3d object recognition and pose. *PAMI*, 1991. 1

[13] A. Frome, D. Huber, R. Kolluri, T. Bülow, and J. Malik. Recognizing objects in range data using regional point descriptors. In *ECCV*. 2004. 1

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1, 6

[15] J. M. Gonfaus, X. Boix, J. van de Weijer, A. D. Bagdanov, J. Serrat, and J. Gonzàlez. Harmony potentials for joint classification and segmentation. In *CVPR*, 2010. 1

[16] A. Gupta, A. A. Efros, and M. Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *ECCV*. 2010. 1

[17] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *CVPR*, 2011. 1

[18] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *CVPR*, 2013. 2, 7, 8

[19] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *ECCV*. 2014. 7

[20] V. Hedau, D. Hoiem, and D. Forsyth. Recovering free space of indoor scenes from a single image. In *CVPR*, 2012. 1

[21] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1), 2007. 1

[22] D. Hoiem, A. Stein, A. A. Efros, and M. Hebert. Recovering occlusion boundaries from a single image. In *ICCV*, 2007. 1

[23] A. Ion, J. Carreira, and C. Sminchisescu. Probabilistic joint segmentation and labeling. In *NIPS*, 2011. 1

[24] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *PAMI*, 21(5), 1999. 1, 2, 4

[25] P. Kohli, L. Ladicky, and P. Torr. Robust higher order potentials for enforcing label consistency. In *CVPR*, 2008. 1

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 6

[27] L. Ladicky, P. Torr, and P. Kohli. Robust higher order potentials for enforcing label consistency. In *CVPR*, 2008. 1

[28] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *International Conference on Robotics and Automation (ICRA)*, 2011. 2

[29] 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. 1

[30] A. Leonardis, A. Jaklic, and F. Solina. Superquadrics for segmenting and modeling range data. *PAMI*, 19(11), 1997. 1

[31] M. Leordeanu, R. Sukthankar, and C. Sminchisescu. Efficient Closed-Form Solution to Generalized Boundary Detection. In *ECCV*, 2012. 1, 3

[32] A. Levinshtein, C. Sminchisescu, and S. Dickinson. Optimal contour closure by superpixel grouping. In *ECCV*, 2010. 1

[33] F. Li, G. Lebanon, and C. Sminchisescu. Chebyshev Approximations to the Histogram χ^2 Kernel. In *CVPR*, 2012. 1

[34] D. Lin, S. Fidler, and R. Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *ICCV*, 2013. 2

[35] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2), 2004. 2

[36] M. Maire, P. Arbelaez, C. Fowlkes, and J. Malik. Using contours to detect and localize junctions in natural images. In *CVPR*, 2008. 1, 3

[37] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010. 1

[38] S. Ramalingam, P. Kohli, K. Alahari, and P. Torr. Exact inference in multi-label crfs with higher order cliques. In *CVPR*, 2008. 1

[39] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *CVPR*, 2012. 1, 2, 7

[40] A. Roy and S. Todorovic. Scene labeling using beam search under mutex constraints. In *CVPR*, 2014. 2

[41] A. Saxena, S. H. Chung, and A. Y. Ng. 3-d depth reconstruction from a single still image. *IJCV*, 76(1), 2008. 1

[42] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1), 2013. 2

[43] N. Silberman and R. Fergus. Indoor scene segmentation using a structured light sensor. In *ICCV Workshops*, 2011. 2

[44] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1, 2, 3, 6, 7, 8

[45] J. Tang, S. Miller, A. Singh, and P. Abbeel. A textured object recognition pipeline for color and depth image data. In *International Conference on Robotics and Automation (ICRA)*, 2012. 2

[46] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *CVPR*, 2010. 1

[47] W. Xia, Z. Song, J. Feng, L.-F. Cheong, and S. Yan. Segmentation over detection by coupled global and local sparse representations. In *ECCV*, 2012. 1

[48] X. Xiong, D. Munoz, J. A. Bagnell, and M. Hebert. 3-d scene analysis via sequenced predictions over points and regions. In *International Conference on Robotics and Automation (ICRA)*, 2011. 2

[49] A. Zisserman, D. Forsyth, J. Mundy, C. Rothwell, J. Liu, and N. Pillow. 3d object recognition using invariance. *Artificial Intelligence*, 1995. 1