Hierarchical Image Segmentation Ensemble for Objectness in RGB-D Images

Huiqun Wang, Di Huang, *Member, IEEE*, Kui Jia, and Yunhong Wang, *Senior Member, IEEE*

***Abstract*— Objectness has recently become a standard step in many computer vision tasks. Among various techniques, those based on hierarchical image segmentation play a fundamental role for developments in new data modalities. In this paper, we address the problem of objectness in RGB-D images and propose a novel and effective approach, namely, hierarchical image segmentation ensemble (HISE). Different from existing image segmentation based methods that generate object segments or proposals largely by heuristics or empirical rules, HISE learns superpixel mergings with a hierarchical tree-structured ensemble, where individual merging models of the ensemble are formed by traversing different paths of the tree, and where both the merging accuracy and proposal diversity are emphasized. Furthermore, we use efficient feature measurements that support easy integration of additional clues. Extensive experiments conducted on the benchmark NYU-v2 RGB-D and SUN RGB-D data sets show the competency of our proposed method.**

***Index Terms*— Object proposal, RGB-D data, segmentation.**

I. Introduction

OBJECTNESS (also known as object proposal generation) aims to sample a reasonable number of local regions (typically less than a few thousands either with precise boundaries or as bounding boxes) from a given image, which are expected to contain all the possible generic object instances in the image. It has received extensive attention in the field of computer vision and pattern recognition in recent years, since it substantially improves the performance and reduces the time consumption of many sophisticated tasks, e.g. object detection, object recognition, and object retrieval.

The pioneering research discusses the object proposal problem in the viewpoint of a traditional sliding window based object detection [1], and it makes use of different appearance and geometry characteristics conveyed in an image window to compute a score, measuring the possibility whether there exists a specific object. Since then the techniques on such an issue have been developed in two main categories, with the same goal to reduce the number of proposed candidates while maintaining a level of recall of the ground-truth object instances in images. One follows the way in [1] but extends it by adopting more effective features [2]–[5] and a learning phase to rank the candidates. Another treats this problem as in image segmentation, and candidates are strategically hypothesized from hierarchical image segmentation results [6]–[12], where superpixels and intermediately merged local regions are natural choices. In either case, there is a concern of what criteria to use so as to discard/rank down those less relevant hypotheses. Properties of hypothesized local regions relevant to "objectness" [13] include size and location, shape, contour features.

More recently, due to the advent of low cost and portable commercial 3-D imaging sensors, e.g. Microsoft Kinect and ASUS Xtion, RGB-D images have become available in many computer vision related tasks. Although the depth cue provided has a low resolution and a limited distance range, it indeed results in improved performance compared to that of only 2-D (RGB) images in indoor scenes. Regarding object proposal, it has been shown that the joint use of the information in both the RGB and D channels outperforms either of the single ones [14]–[17]. These attempts basically expand segment-based methods previously proposed for RGB images to integrate the contributions in the modality of RGB and D. However, current hierarchically merging strategies in RGB images are not sufficiently stable, and the limitation remains in such direct expansion, thus making it problematic when fusing the clues in RGB-D images.

Specifically, segment-based object proposal approaches are usually realized by a bottom-up process which hierarchically merges spatially neighboring superpixels or intermediately generates local regions supposed to possess similar properties of homogeneity. Nevertheless, the criteria of region homogeneity are not always relevant to the ultimate goal of semantic object segmentation, resulting in two common shortcomings: 1) there are too many false positives (merged regions that do not have large overlaps with the ground-truth object instances in images) generated during the hierarchical merging process, i.e., *the precision is low*, and 2) some of the true positives (regions of the ground-truth object instances in images) are not generated by the end of the hierarchical merging process, i.e., *the recall is low*.

Manuscript received July 25, 2017; revised October 23, 2017; accepted November 7, 2017. Date of publication November 21, 2017; date of current version January 7, 2019. This work was supported in part by the National Key Research and Development Plan under Grant 2016YFC0801002 and in part by the Research Program of State Key Laboratory of Software Development Environment under Grant SKLSDE-2017ZX-07. This paper was recommended by Associate Editor J. M. Martinez. *(Corresponding author: Di Huang.)*

H. Wang, D. Huang, and Y. Wang are with Beijing Advanced Innovation Center for Big Data and Brain Computing, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: huiqunwang@buaa.edu.cn; dhuang@buaa.edu.cn; yhwang@buaa.edu.cn).

K. Jia is with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510630, China (e-mail: kuijia@scut.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2017.2776220

For remedy, one may choose to use some additional features [29], [30] in the hierarchical merging process that characterize properties of generic objects and are normally complementary to region homogeneity. But use of these features could do harm to semantic object segmentation since regions that are discarded to be formed by merging, when described by these features, may later be used as components to form a true region of generic object. Alternatively, the ranking step can be added after the hierarchical merging process, which uses features of generic objects and desires to rank highly some of the merged regions that are believed to be the true positives. Both the solutions might somehow improve the precision of generating object segments/proposals from images; unfortunately, they generally cannot improve the recall. Instead, our idea is to sacrifice a bit the objective of merging homogeneous areas, for a more diversified generation of region hypotheses of generic objects, so that as many regions of ground-truth object instances in images as possible can be generated in the hierarchical merging process, i.e. to improve recall while maintaining precision.

This paper proposes a novel and effective Hierarchical Image Segmentation Ensemble (HISE) for object proposal in RGB-D images. HISE is a hierarchical tree-structured superpixel merging ensemble, where individual merging models are formed by traversing different paths of the tree. Each path/model contains sequentially learned classifiers that determine whether two adjacent superpixels/regions should be merged, where superpixels may be intermediate ones generated in this merging process. To achieve proposal diversity, we learn at each (root and internal) node of the tree complementary children classifiers by reducing correlations of their model parameters. To improve merging accuracy for each classifier, we propose soft labeling of adjacent superpixels, which can characterize mergings at object boundaries more precisely. In addition, we use feature measurements that support efficient feature computation for intermediately generated superpixels. Note that existing object proposal techniques are to some extent based on this diversified quality search strategy [3], [6], [7], [12], [16], [18]; however, their implementations of this strategy are only based on heuristics or empirical rules. In contrast, HISE achieves proposal diversity, merging accuracy, and computational efficiency in a principled and systematic way. Extensive experiments on the NYU-v2 RGB-D and SUN RGB-D datasets demonstrate the efficacy of HISE. In particular, we outperform the state of the art methods, including Selective Search [7], EdgeBox [5], MCG [10], etc.

We summarize our technical contributions as follows.

- We propose HISE, a hierarchical tree-structured superpixel merging ensemble, for object proposal in RGB-D images. HISE achieves diversified superpixel generation by learning at each node of the tree children classifiers that are enforced to be less correlated. Consequently, individual models formed by traversing different paths of this learned tree are complementary in merging superpixels and generating proposals.
- To improve accuracies of merging superpixels, especially for those residing at object boundaries, we propose soft labeling of adjacent superpixels to help resolve the merging ambiguities in training samples. Experiments show that soft labeling clearly improves merging accuracy and proposal generation.
- We design a set of histogram based features for learning of merging classifiers. Our features capture both the appearance and geometrical clues in RGB-D images. They also support efficient feature computation for intermediately generated superpixels.

II. Related Work

As described in Section I, there are two major types of approaches to generate object proposals: i.e. window scoring based as well as segment based. In this section, we briefly review the two categories (referring to [19] for a recent survey with more details). In addition, we also give separate summaries of objectness for RGB-D images and that using deep learning techniques.

Window scoring based methods exploit the well known "sliding window" paradigm, which generate a large pool of hypotheses and score each candidate window according to how likely it is to contain an object. Its key procedure lies in what features are used to rank these candidates. The original study of object proposal falls into this category, where Alexe *et al.* [1] initially sample a set of image boxes from salient locations and then rank them based on a combination of multiple low-level features, including color, edge, size, and superpixel straddling. Rahtu *et al.* [2] start with the proposal pool produced from single superpixels and their pairs and triplets, along with some randomly filtered windows. The scoring step in [1] is enhanced by some features of superpixel boundary integral and boundary edge distribution whose optimal combination is learned in a cascade structure. Cheng *et al.* [4] present a fast approach (about 300 fps) using the cue of contour from a sliding window encoded in the norm of gradients at various scales and a simple linear classifier. Zitnick and Dollar [5] propose EdgeBoxes, which begins with a coarse sliding window as well but builds on object boundary estimates through structured decision forests. In addition, they apply a refinement step of greedy iterative local search after initial scoring to improve localization.

Segment based methods produce multiple regions likely corresponding to objects by diversifying hierarchies of image segmentation, and determine whether a hypothesis, i.e. a superpixel or an intermediate merged region, is a candidate when it is created. The principle focuses on the way to combine segments or proposals where their similarity is measured by a set of complementary low level clues. SelectiveSearch (SS) [7], [20] is a typical representative of this category. It greedily merges manually designed superpixels to generate proposals, and has been widely adopted in object detection. Manen *et al.* [8] introduce the superpixel connectivity graph to build random partial spanning trees, and it is a randomized superpixel merging process to learn all probabilities, which achieves significant speed-up. This graph cut idea for foreground-background segmentation also appears in [3], [9], [21], [22]; they use various strategies

for seed generation and some of them apply the ranking step. Krahenbuhl and Koltun [11] display Geodesic Object Proposals. It first generates over-segmentation by using the fast edge detector and then utilizes classifiers to place seeds for a geodesic distance transform. Level sets of each distance transform segment figures from background as proposals. Similarly, Arbeláez *et al.* [10] calculate multi-scale hierarchies of the fast edge detector based over-segmentation; and regions are merged according to edge strength and the resulting object hypotheses are further ranked using common basic features, known as multiscale combinatorial grouping (MCG). Carreira and Sminchisescu [6], [23] propose constrained parametric min-cuts (CPMC), which can be regarded as a special case in this category. It computes graph cuts with different seeds and unaries clustering directly on pixels rather than hierarchically merging initial superpixels. The segments are then ranked using a large pool of features as in window scoring proposal techniques.

Comparing the two types of methods, segment based ones tend to reach higher recall rates, and window scoring based ones generally run faster. but the results of the latter are often with relatively low precision in localization unless the regions are collected very densely [19], which requires subsequent optimization as in [5] and [25].

In the case of object proposals for RGB-D images, the existing studies mostly derive from the ones of RGB images aforementioned. For example, Lin *et al.* [15] extend the CPMC framework to generate candidate cuboids, and develop a conditional random field to integrate information from different sources to classify the cuboids. Gupta *et al.* [14] generalize gPb-UCM hierarchical segmentation [25] to 3D features such as the shape, size, and geocentric pose for better bottom-up segmentation and region proposal. Gupta *et al.* [16] later advance MCG to fuse the information in the RGB and D channels, and a geocentric embedding is proposed to encode height above ground and angle with gravity for each pixel from depth images in addition to the horizontal disparity, which proves better than the raw depth data for representation. In [17], Deng *et al.* develop an unsupervised framework to generate bottom up class independent object candidates. Instance regions are produced by multi-channel multi-scale segmentations in the RGB image and bounding boxes are created according to five different plane based cues in the depth image, where a revised GrabCut is applied to dynamically model global object and background properties.

Additionally, for the great advance achieved in end-to-end model based object detection and recognition, e.g. [26] and [27], there exist a few investigations which apply deep learning techniques. Pinheiro *et al.* [28] propose an approach based on a discriminative convolution neural network which is optimized with respect to the objectives of class-agnostic segmentation mask and likelihood of the patch centered on a full object. It reaches significant performance gain on the databases of PASCAL VOC and MS COCO [31], compared with the hand-crafted counterparts. Pinheiro *et al.* [32] then refine the mask encoding in a top-down pass utilizing features at successively lower layers of the deep network to produce a more reliable segmentation, improving both the accuracy and efficiency.

Our approach, HISE, shares the idea of the multi-branch cascade structure and the automatic learning of complementary merging strategy with [12]. Note that differences lie in the following three major aspects. 1) In contrast to the binary tree structure, HISE presents a multi-stage branching framework. Each stage has a flexible number of branches, which enables us to apply more branches in the top layer and fewer branches in the bottom layer. The structure can thus be deepened without suffering exponentially increased computational cost as in [12]. 2) Instead of increasing the loss weights of wrongly classified samples in a Boosting-like procedure, HISE directly minimizes the correlation between each branch for the purpose of making them complementary. The diversity of segments can hence be expended conveniently when more features are embedded. 3) Unlike the scenario of objectness in RGB data, HISE concentrates on RGB-D images, where disparity features are employed. Thanks to these properties, HISE achieves very competitive results in the given task, which are even superior to the ones of deep model based approaches.

III. HISE Model

The process of superpixel merging in image segmentation commonly follows a bottom-up greedy strategy. In this process, whether to merge a pair of adjacent superpixels is based on some measure of homogeneity, with the pair of highest homogeneity to be merged first. The choice of homogeneity measure thus plays an essential role for better segmentation results. However, as stated in Section I, general criteria of homogeneity may not be always relevant to semantic object segmentation. If an incorrect merging occurs during the process, it would be carried on to the final result. To address this issue, we aim to learn classifiers and use scores of classifiers as the measurement of homogeneity, similar to [33]. The learned classifiers take as input a pair of superpixels and decide whether to merge them or not. Since feature statistics change as larger patches (resulting from superpixel merging) appear in the merging process, we consider a multi-stage strategy where classifiers are sequentially learned at each stage. Since there are no well defined objectives that can measure the geometrical and appearance features of a generic object, we are thus tempted to diversify our criteria by training an ensemble of complementary classifiers.

More specifically, we adopt a hierarchical tree-structured ensemble as illustrated in Fig. 1. At each root or internal node of the tree, we learn complementary children classifiers as described shortly. Each of these classifiers produces a new internal node where superpixels of finer granularity have been merged to a coarser level. We then learn the next-stage complementary children classifiers at each of the resulting nodes. The process continues until a specified number of stages. In this process, we control the ratio of the number of superpixels to be merged at each stage by merging pace (denoted as B in Algorithm 1 and Algorithm 2). When learning is done, individual models of the ensemble are formed by traversing different paths of the tree. In testing, each node in the

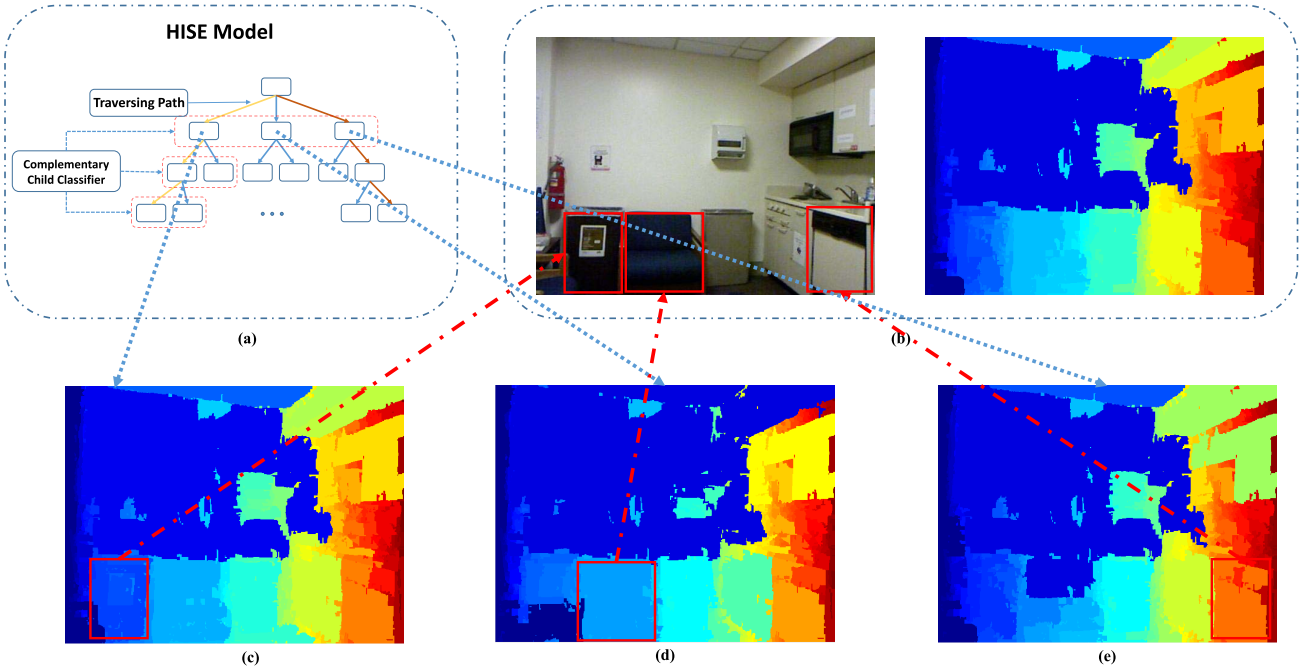

Fig. 1. Framework of the proposed Hierarchical Image Segmentation Ensemble (HISE). (a) Tree structure of HISE, where the Complementary Child Classifier is a group of linear classifiers enforced to minimize the correlation of their model parameters (the training progress is illustrated in Sec III-A); (b) Original image and its corresponding superpixels; and (c), (d), and (e) Results generated by different complementary classifiers. The red boxes show the final prediction of HISE.

Algorithm 1 Object Proposal by HISE

Require:
The initial superpixel set S_0
The classifier of each model $\mathbf{W} = \{\mathbf{w_{t,k}}\}, \mathbf{t} \subset [1, T]$
Merging pace $B = \{b_1, b_2, \ldots b_T\}$ (it controls the number of the adjacent superpixels to be merged).

Initialize: proposal region set:$R = \{S_0\}$
for stage $\mathbf{t} = 1$ to T **do**
  for each node c in stage $\mathbf{t}$ **do**
    load the parents nodes superpixel $S_{\mathbf{t-1},\hat{c}}$
    $(R_{\mathbf{t},1}, \ldots, R_{\mathbf{t},k}) \Leftarrow$ greedy merging $S_{\mathbf{t-1},\hat{c}}$ via $(\mathbf{w_{t,k}}, b_t)$
    add $(R_{\mathbf{t},\mathbf{k_1}}, \ldots, R_{\mathbf{t},\mathbf{k_n}})$ to R
  end for
end for
get bounding box set BB from Region Set R
refine bounding box set BB
return BB

Algorithm 2 The Training Progress for HISE

Require:
The training image set I_{train}
The stage number T
The branch degree for each stage of $\{\mathbf{K_1}, \mathbf{K_2}, \ldots, \mathbf{K_T}\}$
Merging pace $B = \{b_1, b_2, ..b_T\}$

Initialize: generate the original superpixel S_0 from I_{train}
for $\mathbf{t} = 1$ to T **do**
  for each node c in stage $\mathbf{t}$ **do**
    initial $\mathbf{w_{t,c}}$ randomly
    collect training samples $\{\mathbf{f}_{\mathbf{t},c}, l_{\mathbf{t},c}\}$ from parent superpixel $S_{\mathbf{t-1},\hat{c}}$
    train $\mathbf{w}_{\mathbf{t},c}$ with samples $\{\mathbf{f}_{\mathbf{t},c}, l_{\mathbf{t},c}\}$ by SGD
    generate new superpixel by $\{\mathbf{f}_{\mathbf{t},c}, \mathbf{w}_{\mathbf{t},c}, b_\mathbf{t}\}$ for child node
  end for
  add $\mathbf{w}_{\mathbf{t},c}$ to $\mathbf{W_t}$
end for
get classifiers $\{\mathbf{W_t}\}_{\mathbf{t}=1}^{T}$
return $\{\mathbf{W_t}\}_{\mathbf{t}=1}^{T}$

traversing path decides which pair of adjacent superpixels/regions should be merged. It then passes the merging result to its child node. After going through the whole HISE model, we collect the merging result from each node and remove the duplicate proposals as output.

A. Complementary Training of Classifiers

We present in this section how children classifiers are learned at each root or internal node of HISE. Denote the feature vector of each superpixel as $\hat{\mathbf{f}} \in \mathbb{R}^C$. For any pair of superpixels $\hat{\mathbf{f}}$ and $\hat{\mathbf{g}}$, we also compute its channel-wise χ^2 distance, resulting in the feature $\mathbf{f} \in \mathbb{R}^C$ for training of merging classifiers (the details of the color and depth features used are presented in Section III-C). Suppose that we have N samples of training superpixel pairs $\{\mathbf{f}_i, l_i\}_{i=1}^{N}$, where $l_i \in \{-1, 1\}$ is the label of each pair indicating whether they should be merged. To promote diversified merging criteria, we propose in this work the following training objective to

learn multiple classifiers at any root or internal node of HISE.

$$\mathcal{L}_1 = \min_{\mathbf{W}} \sum_{i=1}^{N} Loss(\{\mathbf{f}_i, l_i\}, \mathbf{W}) + \lambda \|\mathbf{W}^T\mathbf{W} - \alpha \mathbf{I}_K\|_F^2 \quad (1)$$

where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K]$ are the K linear classifiers to be learned, λ is a scalar parameter, and $\mathbf{I}_K$ is a diagonal matrix of order K. We are concerned with binary classification, and the $Loss(\cdot)$ in (1) can be specified as either hinge loss or cross-entropy loss. The second term in (1) is introduced to promote incoherence between the learned linear classifiers [34], thus to achieve merging diversity, where α is a scalar scaling parameter. In case of hinge loss, the objective (1) can be specified as follows:

$$\mathcal{L}_2 = \min_{\mathbf{W}} \sum_{i=1}^{N} \max\{\mathbf{0}, \mathbf{1} - l_i \mathbf{f}_i^T \mathbf{W}\}\mathbf{1}_T + \lambda \|\mathbf{W}^T\mathbf{W} - \alpha \mathbf{I}_K\|_F^2. \quad (2)$$

The above problem can be efficiently solved via stochastic gradient descent.

B. Soft Labeling of Adjacent Superpixels

During the training progress, we assign a label to each superpixel, determined by the max intersection area with a ground truth object. But a superpixel generally has intersections with several objects, and if we assign only one label to a superpixel, it would be hard to choose an appropriate threshold for merging. If the threshold is too high, it tends to filter a lot of training samples, and if it is too low, more false samples occur, which increases the error in merging.

To deal with this issue, we change hard labeling (i.e. *−1 or 1*) to soft labeling. For a superpixel that has intersections with several ground truth objects, its soft label is defined as $\hat{\mathbf{p}} = [\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n]^\top$, where $\hat{p}_i$ represents the possibility that the superpixel belongs to object i. For a pair of adjacent superpixels with the ground truth labels $\hat{\mathbf{p}}^1$ and $\hat{\mathbf{p}}^2$, the soft label for their merging is defined as $p = \hat{\mathbf{p}}^{1\top}\hat{\mathbf{p}}^2$. The soft labeling of adjacent superpixels can be naturally used when loss function of (1) is specified as cross-entropy loss, for which we have

$$\mathcal{L}_3 = \min_{\mathbf{W}} \sum_{k=1}^{K}\sum_{i=1}^{N} -p_i \log q_i(\mathbf{w}_k) - (1 - p_i)\log(1 - q_i(\mathbf{w}_k)) + \lambda \|\mathbf{W}^T\mathbf{W} - \alpha \mathbf{I}_K\|_F^2, \quad (3)$$

where $q_i(\mathbf{w}_k) = 1/(1 + \exp(-\mathbf{w}_k \mathbf{f}_i))$ is the estimated merging probability for sample pair i using classifier k. The above problem can be efficiently solved using stochastic gradient descent.

C. Efficient Similarity Measurement

To measure the similarity between superpixels, a set of features is extracted from the RGB and D modality respectively. Combining multiple features enlarges the searching space when merging superpixels, and basically benefits the segmentation precision. But on the other hand, it incurs the increase in computational cost. To make our method operate efficiently, we make use of 15 fast-to-compute features: 8 for 2D images and 7 for depth images. Among them, 12 features are histogram based, calculated from a superpixel itself, and they need to be generated on the superpixels of initial over-segmentation only once. If a region is merged by a number of superpixels, such features can be directly inherited from these superpixels, thus accelerating both the training and testing procedures. Another three features are computed between the adjacent superpixels.

We categorize the 15 features used into 7 groups.

1) Color Channel: Instead of RGB, we extract the feature from the LAB color space. For superpixel s, we split each channel into 32 bins, and calculate the histogram of each channel, termed Cl_s, Ca_s, Cb_s, as three color features.

2) Depth Channel: Traditional depth features, e.g., plane segmentation based and geocentric pose based ones always consume high cost in computation. In contrast, we treat the depth map in a simpler way and the feature produced is similar to the ones in the color channel. Each depth map is first normalized into the range of [0, 1] and split into 32 bins. A 32-bin histogram of each superpixel is then calculated, denoted as De_s.

3) Texture Channel: The pattern of textons is an important cue to capture the characteristics of superpixels. Two types of texture features are adopted. One is convolved with eight oriented even and odd symmetric Gaussian derivative filters and a center surround (difference of Gaussians) filter, and each pixel is associated with a vector 17-d of responses. We then cluster all the responses into 32 bins by K-means as a histogram Tf_s, see [25] for more details. Another is standard SIFT, the original 128-d descriptor is also clustered into 32 bins, denoted as Ts_s.

4) Pointcloud Channel: The pointcloud represents the exact position of an object in a scene. We compute its distribution from x-axis, y-axis, and z-axis, and split each channel into 16-bin histograms, termed CX_s, CY_s, CZ_s, as three pointclould features.

5) Normal Channel: To better encode the shape characteristics, we also calculate the normal values of all the vertices on the depth map, and project them to x-, y-, and z-direction. Their histograms are normalized into 16 bins as three gradient features, denoted as GX_s, GY_s, GZ_s.

The differences between the 12 histogram features, i.e., $Cl_s, Ca_s, Cb_s, De_s, Tf_s, CX_s, CY_s, CZ_s, GX_s, GY_s, GZ_s$, and Ts_s, are measured by using χ^2 distance as

$$H_f(s_a, s_b) = \chi^2(f_{s_a}, f_{s_b}) = \frac{1}{2}\sum_{i=1}^{n} \frac{(f_{s_a}(i) - f_{s_b}(i))^2}{f_{s_a}(i) + f_{s_b}(i)} \quad (4)$$

where n is the dimensionality of the histogram.

6) Size Channel: We also employ the size feature and the fill feature in proposed in SelectiveSearch [7], which are both computed between neighboring superpixels. The size feature encourages the smaller superpixel to merge earlier. Its similarity is defined as the fraction of the image that superpixel s_a and s_b jointly occupy:

$$H_{size}(s_a, s_b) = 1 - \frac{size(s_a) + size(s_b)}{size(I)} \quad (5)$$

where I represents the image.

The fill feature measures how well region s_a and s_b fit into each other. It suggests the superpixels to merge to fill the gaps. If two superpixels hardly touch each other, they likely form a strange region and should not be merged. We define BB_{ab} as the tight bounding box around s_a and s_b. The similarity of the fill feature is the fraction of the image contained in BB_{ab} which is not covered by the regions of s_a and s_b

$$H_{fill}(s_a, s_b) = 1 - \frac{size(BB_{ab}) - size(s_a) - size(s_b)}{size(I)}. \quad (6)$$

7) Edge Channel: In the superpixel merging process, the stronger the edge response is, the higher the possibility that two superpixels should not be merged. So we trace back to the initial edge detection result in [25], and define $H_{edge}(s_a, s_b)$ as the mean strength of the adjacent edge.

The edge similarity is computed as

$$H_{edge}(s_a, s_b) = \sum_{(x_i, y_i) \in l} \frac{M_{edge}(x_i, y_i)}{|l|} \quad (7)$$

where M_{edge} is the edge strength; l represents the connective part of two superpixels; and $|l|$ represents the length of the connective part.

IV. Experimental Results

In order to validate the proposed approach, i.e. HISE, in object proposal for RGB-D images, we conduct extensive experiments on two public databases, namely NYU-v2 RGB-D [35] and SUN RGB-D [36]. The databases, settings, and results are described subsequently.

A. Settings

In the experiments on the two databases, we follow the same protocols as used in the previous literature for fair comparison. Specifically, in NYU-v2 RGB-D, 795 images are used for training and 654 images for testing [16], [17]; in SUN RGB-D, the standard splits are adopted. For the SUN RGB-D, 1924 and 1860 samples compose the training and testing set [17]. The recall score of bounding box proposals is calculated with respect to the ratio of positive predictions that exceed the intersection over union (IoU) of 0.5 and 0.7, over the number of all the ground truth objects. The average recalls (AR) [19] for 10, 100, and 1000 proposals are also adopted to contrast previous deep model based approaches such as DeepMask [28] and SharpMask [32].

In HISE implementation, we employ the technique presented in [25] that applies the watershed algorithm to ultra contour map (UCM) to generate initial over-segmentation, where the contour probability is computed using the edge detection result produced by Edgebox [5] for its computational efficiency. To balance the size of superpixels in different images, we set a flexible threshold in the watershed algorithm, and the number of original superpixels is averagely maintained at 1800. MTSE [24] and EdgeBox [5] are applied to refine and rank the proposals.

In NYU-v2 RGB-D and SUN RGB-D, the 15 features described in Section III-C are exploited. For all the experiments, the number of HISE stages and its branch degree are set at 5 and [4, 3, 3, 2] respectively, which means that in the 5-stage model the first stage has 4 branches and each node at the second, third, and fourth stage has 3, 3, and 2 branches respectively. The merging pace is tuned between 0.15 to 0.2. To increase the orthogonality of individual degrees, we make λ proportional to the degree number with the instant coefficient at 0.5.

For DeepMask and SharpMask, we use the models pre-trained on MS-COCO and Pascal VOC, and select the top 1000 proposals for comparison.

B. Results

1) Performance on NYU-v2 RGB-D: In this experiment, we validate HISE. In comparison, we select three state of the art methods for RGB-D data as counterparts, including MCG-3D [16], CPMC-3D [15] and Deng [17]. In addition, the results of SelectiveSearch (SS) [7], EdgeBox (EB) [5], and MCG [10] achieved only on the RGB modality are displayed as baselines.

In Fig. 2, we can see that HISE delivers very competitive recall rates for different numbers of proposals with IoU at 0.5 and 0.7, which are better than EdgeBox [5], CPMC [15], and Deng [17] on the whole. When we focus on the comparison between MCG3D [16] and HISE, the cases in both sub-figures are similar, where MCG3D works better with smaller numbers of proposals (i.e., less than 1500) while HISE shows its advantage with larger numbers. As the recall rate of a smaller number highly depends on the proposal ranking scheme, which aims to select better candidates earlier, it suggests the one adopted in MCG3D is more powerful. Then, a problem may naturally arise: why not use the same scheme in HISE? Unfortunately, MCG3D ranks proposals based on a rich set of features (43 types) that are previously used in proposal generation and HISE only makes use of 15 basic ones; their direct combination requires calculation of additional features, greatly increasing the complexity. Instead, we use the ranking scheme proposed in EdgeBox [5], which is more convenient to link to HISE and more efficient in computation.

Fig. 3 displays more detailed recalls, where HISE and MCG3D [16] outperform the others. When we compare HISE and MCG3D, for different numbers of proposals as in the three sub-figures, MCG3D reports better recall rates as IoU is larger than 0.7. It indicates its bounding boxes are closer to ground truths, but as we introduce, such performance is achieved by jointly using more than 40 kinds of features. Indeed, combining more features tends to reach better accuracies, which can be clearly evidenced by the performance gain when the features in the D modality are integrated to the ones in RGB in the HISE model. It should also be noted that thanks to its tree structure, HISE possesses the property to integrate additional features in a more flexible way, and it thus has the potential to be ameliorated as more features are used, especially for larger IoU values. At last, recall that the result with IoU at 0.7 is still the most important indicator since accuracy and efficiency are well balanced [5] where HISE gives better scores only based on 15 basic features.

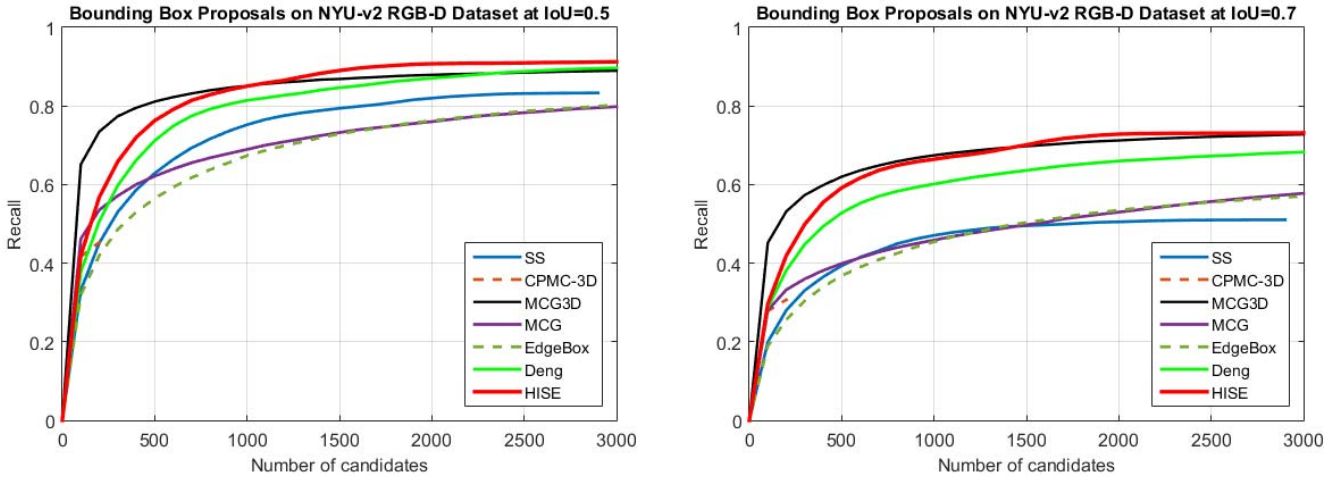

Fig. 2. Comparison of recall with respect to the number of candidates with IoU at 0.5 and 0.7 on the NYU-v2 RGB-D Dataset.

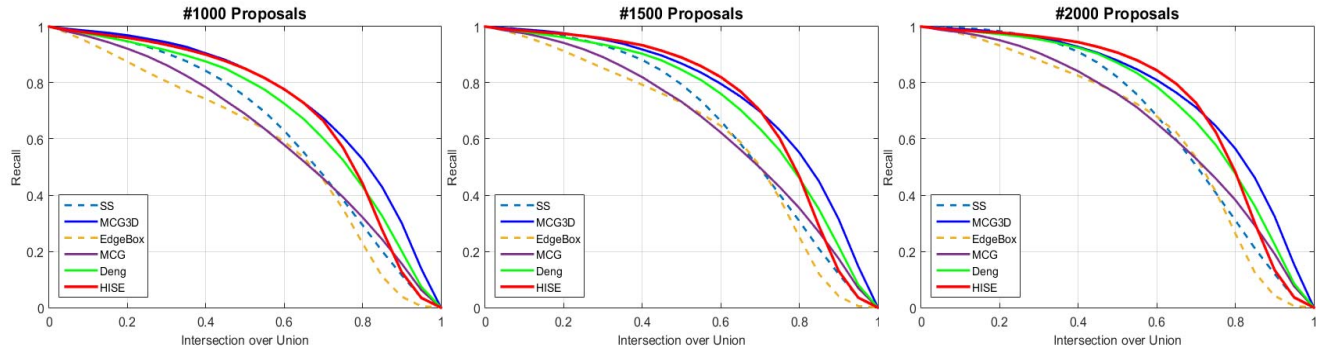

Fig. 3. Recall rates using different IoU thresholds for fixed proposal budgets on the NYU-v2 RGB-D dataset.

2) Comparison With Deep Learning Methods: Deep learning methods deliver good results in many tasks in computer vision (e.g., image classification or object detection). However, the success is largely enabled by the availability of huge annotated domain-specific datasets. Despite RGB-D sensors being sharply increased and widely used, RGB-D data are not as plentiful as RGB images to well support training deep models, which can be illustrated by the fact that current RGB-D datasets are much smaller than those of RGB images. For example, in NYU-D, a popular benchmark to evaluate RGB-D scene analysis methods, the training set only consists of 795 images, based on which it is really hard to build a satisfying deep network.

Compared to deep learning methods, HISE is a lightweight model to generate object proposals on RGB-D images, which only needs thousands of model parameters to reach competitive performance with limited training samples. Conversely, deep models often have more than dozens of millions of parameters. We compare HISE with two popular deep methods, Deep Mask [28] and Sharp Mask [32] for their good performance, and they prove more effective than Region Proposal Network (RPN). We report AR scores at 10, 100, and 1000 proposals, and list detailed recalls of 40 classes at IoU=0.7 with 1000 proposals. From Table I, we can find that deep model based methods tend to reach a higher recall at a small amount of proposals. Although the result of HISE is slightly inferior to those of DeepMask and SharpMask at AR@10, it does not make much sense, since the recall rates are too low to support real applications. With 100 proposals, HISE reaches a comparable result with them. With 1000 proposals, the result is superior to theirs.

TABLE I
COMPARISON WITH DEEP MODELS IN TERMS OF AR AT 10, 100, AND 1000 PROPOSALS ON THE NYU-v2 RGB-D DATASET

	AR@10	AR@100	AR@1000
HISE-RGBD	0.058	**0.246**	**0.529**
HISE-RGB	0.043	0.209	0.452
DeepMask[29]	0.089	0.246	0.417
SharpMask[33]	**0.091**	0.245	0.416

In the meantime, we analyze the results at class level as in Fig. 4. We divide the 894 classes into 40 coarse groups following the definition in [14] and then calculate the class-wise recall with 1000 proposals at 0.7 IoU. As shown in Fig. 4, we can see that HISE reports comparable result with DeepMask and SharpMask on the classes appeared in Pascal VOC and MS-COCO, e.g., *television* and *chair*. However, both datasets provide many more samples in each class than NYU-v2 RGB-D does. Regarding the other classes,

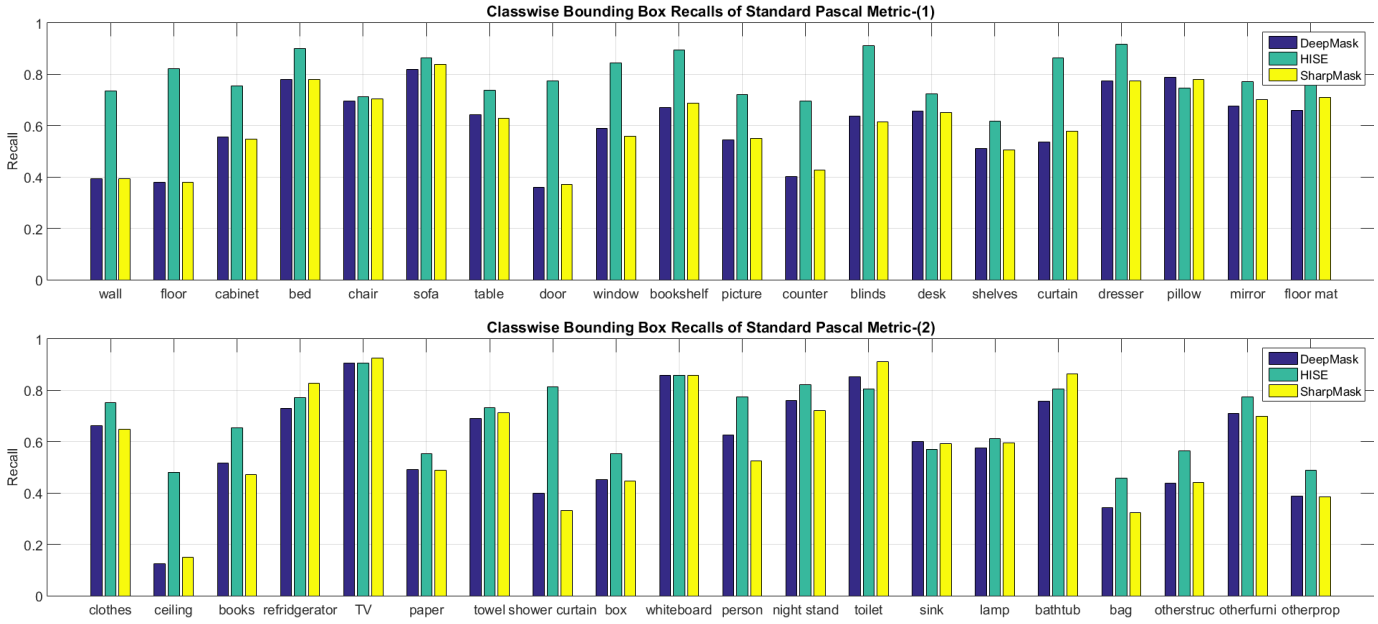

Fig. 4. Class-wise recalls of HISE, DeepMask and SharpMask with IoU at 0.7 and 1000 proposal on the NYU-v2 RGB-D dataset.

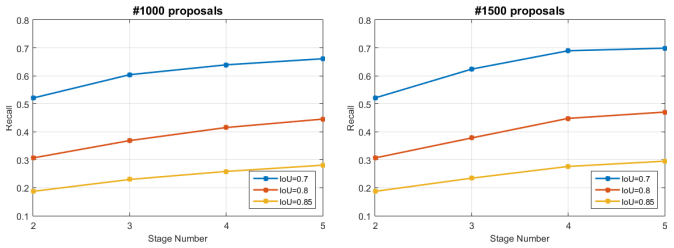

Fig. 5. Recall with different amount of stage at IoU=0.7 on the NYU-v2 RGB-D Dataset.

TABLE II
AR SCORES OF HARD-LABEL AND SOFT-LABEL ON THE NYU-V2 RGB-D DATASET

	Soft-Label	Hard-Label
500	**0.443**	0.439
1000	**0.516**	0.489
1500	**0.541**	0503
2000	**0.553**	0.526

HISE performs significantly better than they do. In particular, for the last three classes (other furniture, structure and property), which include 68, 82, and 707 subclasses respectively and are unseen to all the methods, the performance of HISE is superior to those of deep models, which proves its generality. Notice that HISE is trained on only 798 images while DeepMask and SharpMask are trained on two large-scale benchmarks.

3) Model Analysis: In Fig. 5, we report the recall rate with a fixed amount of proposals. We can find that HISE achieves better results with more stages. In HISE, branches are forced to be less correlated, resulting in diverse proposal generation. We report the class-wise recall at IoU of 0.7 with 2000 proposals in Fig. 6. We can see that each branch has its own contribution, and when HISE combines them, the best performance is reached. We also compare the results of Soft-Label and Hard-Label with a 7-stage model, where each node has two complementary children. AR with the amount of proposals is reported in Table II, and it can be seen that Soft-Label substantially improves the Hard-Label based accuracies.

4) Ablation Study: There are two major parameters that have major impacts on the performance of HISE, i.e., stage number and branch degree. The stage number controls the merging pace while the branch degree corresponds to the diversity of the merging result. Specifically, if we use a model of T stages and each stage has K branches, there will be $(K^T - 1)/(K - 1)$ hierarchical segmentation results. We can see that the computational complexity increases exponentially as the stage number and branch degree grow, and it is thereby necessary to make a proper trade-off between precision and efficiency.

In general, the model of a smaller stage number needs a larger merging pace. If T is fixed, the one with a bigger value of branch degree produces more diverse results. Considering that the number of adjacent superpixels becomes smaller at a later stage, this value should be larger than that at an earlier stage.

In this study, we tune these parameters experimentally. Table III shows such analysis on three models with different branch and stage numbers. As we can see in this table, $T4K3$ outperforms $T3K3$, indicating that the model of a larger stage number reaches a better recall rate. Regarding comparison between $T4K3$ and $T4K2$, it can be seen that based on the same stage number, the model with more branches achieves better performance. For the HISE model in the experiments, it has five stages and the branch numbers of the first four stages are set at 4, 3, 3, and 2, respectively.

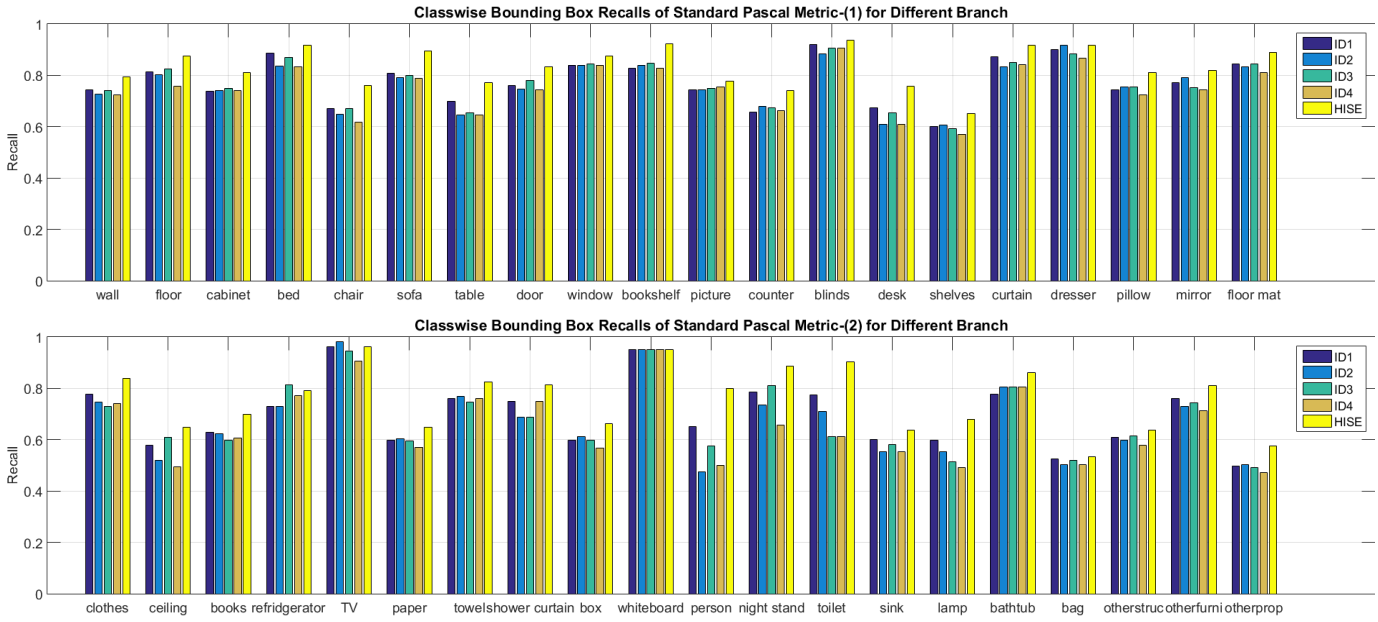

Fig. 6. Recalls of different branch numbers of HISE at IoU=0.7 by 2,000 proposals.

TABLE III

RECALL RATES OF VARIOUS NUMBERS OF PROPOSALS WITH IOU AT 0.7 ON THE NYU-V2 RGB-D DATASET (T DENOTES THE NUMBER OF STAGES; K DENOTES THE NUMBER OF BRANCH DEGREE; AND THE SCORES IN THE BRACKETS ARE THE NUMBERS OF NODES IN THE MODELS, REPRESENTING THEIR MERGING RESULTS)

	$T3K3$(13)	$T4K2$(15)	$T4K3$(40)	HISE-F (125)
500	0.510	0.493	0.527	0.592
1000	0.590	0.586	0.605	0.665
1500	0.605	0.592	0.639	0.701
2000	-	-	-	0.728
2500	-	-	-	0.731

We can see that our model, noted as HISE-F in Table III, reports the best result. Additionally, it should be noted that $T3K3$, $T4K2$ and $T4K3$ do not display the recall rates when the number of proposals is larger than 2000, since such small models cannot produce enough proposals.

5) Performance on SUN RGB-D: Since all the images from the NYU-v2 RGB-D dataset are acquired by Microsoft Kinect v1 which measures the depth cue in a very low accuracy, we choose for validation a subset from SUN RGB-D whose samples are collected through Kinect v2, an advanced version of a relatively high precision; we observe how HISE works when the image quality varies. Meanwhile, we directly compare the result obtained by HISE with that in [17] in terms of recall rate when the number of candidates is set at 3000.

We can see from Table IV that HISE outperforms [17], finding a smaller number of proposal with higher qualities. To be specific, HISE reaches the recall rate of 94.1% and 93.1% as the number of candidates is 2,971 and 2000 respectively, both of which are better than the one of 90.8% provided by [17] at 2,971 candidates. It confirms the advantage of the proposed method over [17] for generating object proposals in RGB-D images.

TABLE IV

COMPARISON BETWEEN HISE AND [17] IN TERMS OF RECALL USING DIFFERENT NUMBERS OF CANDIDATES WITH IOU AT 0.5 ON THE SUBSET IN SUN RGB-D

Sensors	Kinect v2	
Candicate Number	2000	2971
Deng *et al.* [17]	-	0.908
HISE	0.931	**0.941**

TABLE V

RECALL RATES OF HISE WITH VARIOUS NUMBERS OF PROPOSALS AND DIFFERENT IOU THRESHOLDS ON THE SUN RGB-D DATASET

	IoU=0.5	IoU=0.7
500	0.840	0.622
1000	0.899	0.676
1500	0.915	0.695
2000	0.931	0.711

We can see more details in Table V, where HISE is able to achieve a high recall rate with a small set of candidates. For example, HISE obtains a recall of 0.899 with only 1000 proposals when IoU at 0.5, and it is still at a very promising level (i.e. 0.676) when the IoU threshold changes to 0.7.

We evaluate the effectiveness of the orthogonality term by comparing the results of the model given by (3) (denoted as HISE) to a baseline reached by a model without this term (denoted as HISE-NT). The scores of AR on NYU-v2 RGB-D are displayed in Table VI, where we can see that the orthogonality term in our HISE model contributes to the performance gain.

C. Visualization of Results

See Fig. 7 for the object proposals of some typical samples on the NYU-v2 RGB-D database. The blue bounding boxes

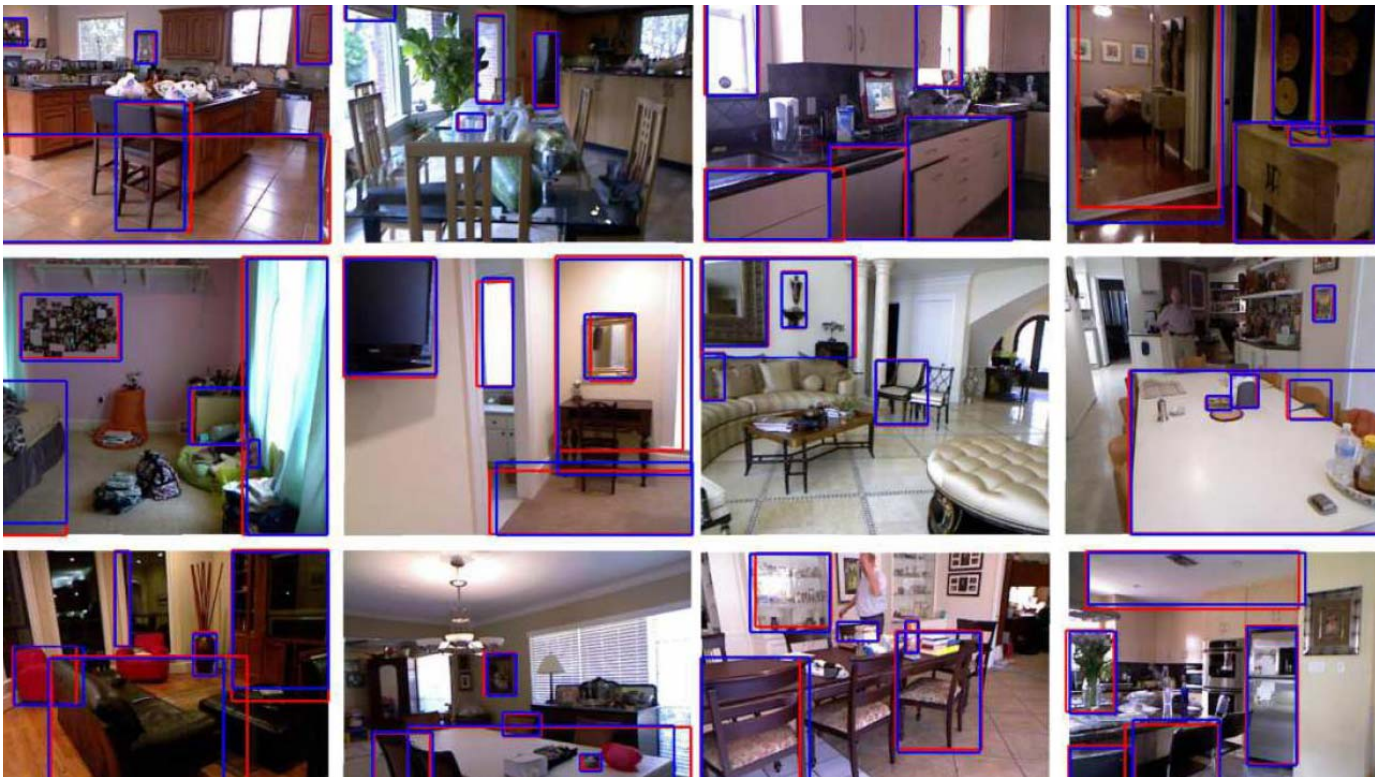

Fig. 7. Demonstration of object proposal results of some typical samples on the NYU-v2 RGB-D database (the blue bounding boxes represent the ground truths and the red ones are the results of HISE).

TABLE VI
AR SCORES OF HISE AND HISE-NT WITH DIFFERENT NUMBERS OF PROPOSALS ON THE NYU-v2 RGB-D DATASET

	HISE	HISE-NT
500	**0.474**	0.352
1000	**0.529**	0.405
1500	**0.556**	0.409
2000	**0.573**	–

represent the ground truths, and the red ones are the results of HISE. We can see that HISE generates very good proposals on the objects of different sizes.

V. CONCLUSION

In this paper, we propose a novel approach, namely Hierarchical Image Segmentation Ensemble (HISE), to object proposal in RGB-D images of indoor scenes. In contrast to current hierarchical segmentation based techniques that achieve a diversified generation of hierarchical image segmentations according to heuristics or empirical rules, HISE extremizes the diversified quality search strategy in a systematic framework of learned tree-structured ensemble, and slightly sacrifices the strength of each classifier by the enforced constraints. Furthermore, due to its structure design, HISE only moderately expands the searching space of the segment composition, and holds a good flexibility as additional features are integrated. We extensively validate it on the NYU-v2 RGB-D and SUN RGB-D databases, the state of the art results achieved demonstrate its competency at finding proposals in RGB-D images.

In the future, we will investigate the way to improve HISE, in particular for the scores of smaller numbers of proposals and the ones under larger IoU thresholds, through advanced ranking schemes and integration of more complementary features in the color and depth channels respectively.

REFERENCES

[1] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *Proc. CVPR*, Jun. 2010, pp. 73–80.
[2] E. Rahtu, J. Kannala, and M. Blaschko, "Learning a category independent object detection cascade," in *Proc. ICCV*, Nov. 2011, pp. 1052–1059.
[3] I. Endres and D. Hoiem, "Category-independent object proposals with diverse ranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 222–234, Feb. 2014.
[4] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300 fps," in *Proc. CVPR*, Jun. 2014, pp. 3286–3293.
[5] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. ECCV*, 2014, pp. 391–405.
[6] J. Carreira and C. Sminchisescu, "CPMC: Automatic object segmentation using constrained parametric min-cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1312–1328, Jul. 2012.
[7] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Apr. 2013.
[8] S. Manen, M. Guillaumin, and L. Van Gool, "Prime object proposals with randomized Prim's algorithm," in *Proc. ICCV*, Dec. 2013, pp. 2536–2543.
[9] P. Rantalankila, J. Kannala, and E. Rahtu, "Generating object segmentation proposals using global and local search," in *Proc. CVPR*, 2014, pp. 2417–2424.
[10] P. Arbeláez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. CVPR*, 2014, pp. 328–335.

[11] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *Proc. ECCV*, 2014, pp. 725–739.
[12] C. Wang, L. Zhao, S. Liang, L. Zhang, J. Jia, and Y. Wei, "Object proposal by multi-branch hierarchical segmentation," in *Proc. CVPR*, Jun. 2015, pp. 3873–3881.
[13] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.
[14] S. Gupta, P. Arbeláez, and J. Malik, "Perceptual organization and recognition of indoor scenes from RGB-D images," in *Proc. CVPR*, Jun. 2013, pp. 564–571.
[15] D. Lin, S. Fidler, and R. Urtasun, "Holistic scene understanding for 3D object detection with RGBD cameras," in *Proc. ICCV*, Dec. 2013, pp. 1417–1424.
[16] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Proc. ECCV*, 2014, pp. 345–360.
[17] Z. Deng, S. Todorovic, and L. J. Latecki, "Unsupervised object region proposals for RGB-D indoor scenes," *Comput. Vis. Image Understand.*, vol. 154, pp. 127–136, Jan. 2016.
[18] J. Pont-Tuset, P. Arbeláez, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 128–140, Jan. 2017.
[19] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 4, pp. 814–830, Apr. 2016.
[20] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders, "Segmentation as selective search for object recognition," in *Proc. ICCV*, Nov. 2011, pp. 1879–1886.
[21] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Proc. ICCV*, Nov. 2011, pp. 914–921.
[22] A. Humayun, F. Li, and J. M. Rehg, "RIGOR: Reusing inference in graph cuts for generating object regions," in *Proc. CVPR*, Jun. 2014, pp. 336–343.
[23] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *Proc. CVPR*, Jun. 2010, pp. 3241–3248.
[24] X. Chen, H. Ma, X. Wang, and Z. Zhao, "Improving object proposals with multi-thresholding straddling expansion," in *Proc. CVPR*, Jun. 2015, pp. 2587–2595.
[25] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, May 2011.
[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 91–99.
[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.
[28] P. O. Pinheiro, R. Collobert, and P. Dollár, "Learning to segment object candidates," in *Proc. NIPS*, 2015, pp. 1990–1998.
[29] X. Liu, J. He, and B. Lang, "Multiple feature kernel hashing for large-scale visual search," *Pattern Recognit.*, vol. 47, no. 2, pp. 748–757, Feb. 2014.
[30] X. Liu, L. Huang, C. Deng, B. Lang, and D. Tao, "Query-adaptive hash code ranking for large-scale multi-view visual search," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4514–4524, Oct. 2016.
[31] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014.
[32] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *Proc. ECCV*, 2016, pp. 75–91.
[33] Z. Ren and G. Shakhnarovich, "Image segmentation by cascaded region agglomeration," in *Proc. CVPR*, Jun. 2013, pp. 2011–2018.
[34] I. Ramírez, F. Lecumberry, and G. Sapiro, "Universal priors for sparse signal modeling: Marrying information theory with sparse coding," Inst. Math. Appl., Univ. Minnesota, Minneapolis, MN, USA, Tech. Rep. 2279, May 2009. [Online]. Available: http://iie.fing.edu.uy/publicaciones/2009/RLS09a
[35] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. ECCV*, 2012, pp. 746–760.
[36] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *Proc. CVPR*, Jun. 2015, pp. 567–576.

Huiqun Wang received the B.S. and M.S. degrees in computer science from Beihang University, Beijing, China, in 2015 and 2018, respectively.

His research interest is object detection.

Di Huang (M'11) received the B.S. and M.S. degrees in computer science from Beihang University, Beijing, China, in 2005 and 2008, respectively, and the Ph.D. degree in computer science from the Ecole Centrale de Lyon, Lyon, France, in 2011.

He then joined as a Faculty Member with the Laboratory of Intelligent Recognition and Image Processing, School of Computer Science and Engineering, Beihang University, where he is currently an Associate Professor. His research interests include biometrics, 2-D/3-D face analysis, image/video processing, and pattern recognition.

Kui Jia received the B.Eng. degree in marine engineering from Northwestern Polytechnical University, China, in 2001, the M.Eng. degree in electrical and computer engineering from the National University of Singapore in 2003, and the Ph.D. degree in computer science from the Queen Mary University of London, London, U.K., in 2007.

He is currently a Professor with the School of Electronic and Information Engineering, South China University of Technology. His research interests are in computer vision, machine learning, and image processing. In these areas, he has authored many publications at prestigious journals and conferences, such as the IEEE T-PAMI, IJCV, T-IP, T-SP, CVPR, ICCV, and ECCV. His recent research focuses on theoretical deep learning and its applications in various computer vision problems, including object recognition, analysis of human activities, and deep learning of 3-D data.

Yunhong Wang (SM'15) received the B.S. degree in electronic engineering from Northwestern Polytechnical University in 1989, the M.S. and Ph.D. degrees in electronic engineering from the Nanjing University of Science and Technology in 1995 and 1998, respectively.

She was with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, from 1998 to 2004. Since 2004, she has been a Professor with the School of Computer Science and Engineering, Beihang University, where she is also the Director of Laboratory of Intelligent Recognition and Image Processing, Beijing Key Laboratory of Digital Media. Her research interests include biometrics, pattern recognition, computer vision, data fusion, and image processing.