Stereo Object Proposals

Shao Huang, *Member, IEEE,* Weiqiang Wang, *Member, IEEE,* Shengfeng He, *Member, IEEE,* and Rynson W.H. Lau, *Senior Member, IEEE*

***Abstract*—Object proposal detection is an effective way of accelerating object recognition. Existing proposal methods are mostly based on detecting object boundaries, which may not be effective for cluttered backgrounds. In this paper, we leverage stereopsis as a robust and effective solution for generating object proposals. We first obtain a set of candidate bounding boxes through adaptive transformation, which fits the bounding boxes tightly to object boundaries detected by rough depth and color information. A two-level hierarchy composed of proposal and cluster levels is then constructed to estimate object locations in an efficient and accurate manner. Three stereo based cues "exactness", "focus" and "distribution" are proposed for objectness estimation. Two-level hierarchical ranking is proposed to accurately obtain ranked object proposals. A stereo dataset with 400 labeled stereo image pairs is constructed to evaluate the performance of the proposed method in both indoor and outdoor scenes. Extensive experimental evaluations show that the proposed stereo based approach achieves better performance than the state-of-the-arts with either a small or a large number of object proposals. As stereopsis can be a complement to the color information, the proposed method can be integrated with existing proposal methods to obtain superior results.**

***Index Terms*—stereopsis, objectness estimation, object proposals, stereo object proposals.**

I. Introduction

OBJECT proposal detection has attracted a lot of attention and made tremendous progress in recent years [1], [2], [3], [4], [5], [6]. Its goal is to create a relatively small set of candidate bounding boxes that cover all the objects in the image. The idea of object proposals is originally motivated by the bottleneck of object detection. Since most state-of-the-art detectors need to exhaustively examine all possible locations and scales in a sliding window fashion [4], [7], [8], [9], [10], resulting in hundreds of thousand of candidate bounding boxes, they are constrained by this computational bottleneck and cannot apply complex models for accurate detection.

Object proposals are typically generated from a large number of candidate windows. Objectness is used to describe the confidence that a window covers an object of any category [1], [2], [11]. The objectness scores of the candidate windows are estimated using different cues, such as surrounding contrast [1], normed gradients [2] and structured edges [11]. By ranking these windows based on the objectness scores, object proposals can then be produced by selecting the top ones from the list. Another way of detecting object proposals takes advantage of multiple overlapping segments to reduce the errors caused by non-overlapping segmentation [12], [13], [14]. The idea of superpixel merging is widely used for generating object proposals with high accuracy [5], [6], [15], [16]. Notwithstanding the demonstrated success, existing RGB-based algorithms may become ineffective when objects cannot be easily separated from the background (e.g., objects with similar colors to the background). In this case, additional cues are required as a complement to detect the objects in the image.

In this work, we leverage stereopsis to overcome the above barrier. As a complement to color images, a stereo image pair is utilized to obtain rough depth and edge correspondences for two images. We first generate a small set of initial bounding boxes, followed by adaptive transformation to fit each box to an object. This transformation is performed based on depth and color information to produce tight object proposals. However, as the transformed bounding boxes tend to cluster around objects, we build a two-level proposal hierarchy, where the cluster level extracts the approximated object locations to reduce redundancy and the proposal level retains the diversity. To estimate the objectness for a given bounding box, our first observation is that different parts of an object usually share similar depth values and are thus on the same depth plane. In addition, photographers tend to arrange objects in a depth level close to the camera. This means that regions with small depth values are more likely to be objects. Further, to handle complex and challenging scenarios, e.g., persons standing shoulder to shoulder or the perspective view of a large object such as a train moving toward the camera, we propose another cue to utilize edge information (as motivated by [1], [11]) to handle complicated objects.

To evaluate the proposed method, we have constructed a new dataset of 400 stereo image pairs. Extensive experimental evaluations show that the proposed method achieves superior performance than all the state-of-the-art methods tested on this new dataset, especially with a small number of proposals. The proposed method can be also treated as a complement of the existing methods. We have integrated the proposed approach into the color-based object proposal methods [1], [2], [6], [11] and obtained performance enhancement on the new dataset.

The main contributions of this work can be summarized as:

- This work computes object proposals from a stereo image pair. Leveraging stereopsis, adaptive transformation is proposed to produce tightly-fit object proposals and a two-level hierarchy is built to handle bounding boxes with similar/dissimilar positions and scales.

Shao Huang is with the School of Computer and Controlling Engineering, University of Chinese Academy of Sciences, CAS, China, and the Department of Computer Science, City University of Hong Kong, Hong Kong. E-mail: shaohuang6-c@my.cityu.edu.hk.

Weiqiang Wang is with the School of Computer and Controlling Engineering, University of Chinese Academy of Sciences, CAS, China. Email: wqwang@ucas.ac.cn

Shengfeng He is with the School of Computer Science and Engineering, South China University of Technology, China. E-mail: shengfeng_he@yahoo.com.

Rynson W.H. Lau is with the Department of Computer Science, City University of Hong Kong, Hong Kong. E-mail: Rynson.Lau@cityu.edu.hk.

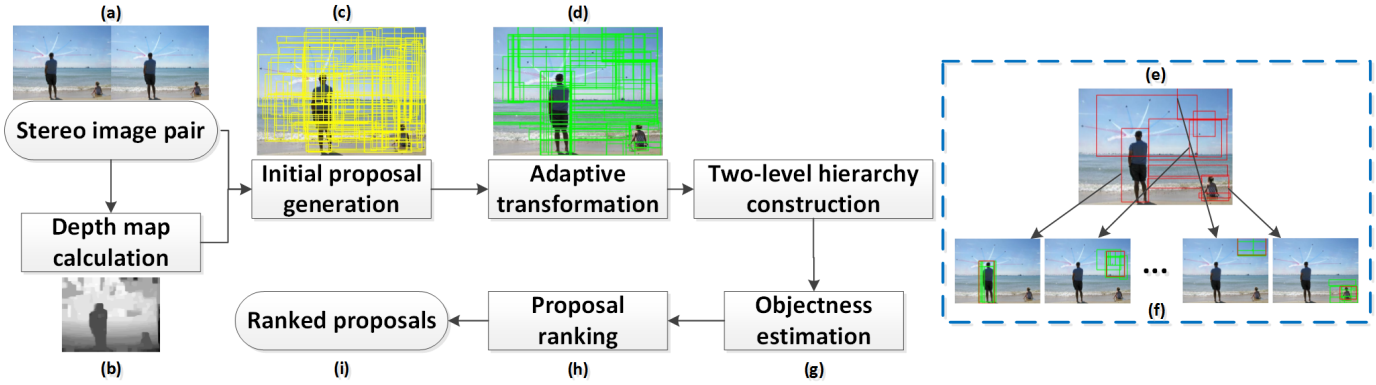


Fig. 1. **The framework of the proposed stereo-based approach.** We start from a stereo image pair and the estimated rough depth map, followed by bounding box initialization and adaptive transformation to generate object proposals. Then, a two-level proposal hierarchy is built. Objectness scores are computed using stereo-based cues. Finally, proposals in the two-level hierarchy are ranked.

- Three stereo based cues "exactness", "focus" and "distribution" are proposed for objectness estimation.
- Color and stereo information are integrated together in different steps to enhance individual inputs. For example, color information is used to enhance the quality of the initial rough depth map; stereo information is used to enhance the performance of existing color-based methods.
- A new stereo dataset with indoor and outdoor scenes is constructed for evaluation. This dataset can be used in other object-related applications of stereo images.

The rest of this paper is organized as follows. Section II describes the related work. Section III presents our stereo-based approach, and Section IV discusses the integration of our approach with RGB-based methods. Section V shows experimental results.

II. Related Work

Object proposal detection aims at finding candidate bounding boxes that cover all objects in the image. Although most of these algorithms are based on color information, some methods make use of other information such as RGB-D. In this section, we review these two types of methods as well as stereo-based segmentation methods, which are related to our work.

Color-based Object Proposals: According to the way that object proposals are computed, object proposal methods can be roughly divided into *window scoring methods* and *superpixel grouping methods*. (See [17] for a comprehensive review.) Given a large set of candidate bounding boxes obtained from different sampling strategies like sliding window, window scoring methods [1], [2], [11] focus on estimating objectness scores of these bounding boxes. The final object proposals are obtained by ranking these boxes to pick the top ones. Different cues are applied to estimate objectness, including superpixels straddling [1], normed gradients [2] and structured edges [18]. The main advantage of these methods is the computational efficiency. However, they are less accurate as they quantize the initial bounding boxes into fixed sizes. Superpixel grouping methods aim at localizing the objects by combining multiple superpixels into object proposals. Carreira *et al.* [19] solve a constrained parametric min-cut (CPMC) problem to produce a set of proposals. Endres *et al.* [20] apply binary foreground-background segmentation on each initialized seed region, and a regressor is trained to predict if a group of segments belong to an object. A learning-free approach, selective search [6], is proposed to generate proposals based on greedy superpixel merging according to low-level features. Due to the use of superpixels, these methods are able to obtain high quality object proposals compared with the window scoring methods. For the same reason, they are typically much slower. Our method falls into the window scoring approach, but we also share some similar properties with the superpixel grouping approach to actively search for objects. We benefit from the stereo image pair by adaptively transforming a relatively small set of initial bounding boxes to tightly fit to the objects. To handle similar bounding boxes to reduce redundancy, we also utilize a clustering algorithm to construct a two-level hierarchy of proposal and cluster levels. As a result, the quality of our proposals outperforms all the other methods tested.

Some recent works introduce a post-processing step to improve proposal quality with a small amount of computation overhead. Long *et al.* [21] propose to use a supervised descent search for re-localizing objects according to the initial proposals. He and Lau [22] present an iterative searching algorithm to locate oriented object proposals. Wang *et al.* [23] improve superpixel-based proposals by using multi-thresholding straddling expansion. The proposed method shares a similar objective with these methods. Unlike them, the proposed two-level hierarchy provides a coarse-to-fine localization process, and thus reduces redundant refinement operations. The added stereo information also leads to an effective scoring function for object localization.

RGBD-based Object Proposals: Due to the difficulty of locating all objects in an image with a small set of proposals, additional information has been introduced for object proposal detection. RGB-D images are the most common alternative input [24], [25], [26]. Bleyer *et al.* [25] show that depth estimation could be improved by introducing the notion of object-level color models as a soft constraint. Bertamini [24] explores how sensitivity to reflection and translation can be modulated by objectness with depth information. Benefiting

from the CNN architecture, Gupta *et al.* [26] propose a geocentric embedding approach for depth images to encode height above ground and angle with gravity for each pixel to construct depth features. In robotics, huge progress has been made on RGB-D integration to guide robot movement in indoor environments [27], [28], [29], [30]. A novel joint optimization algorithm [27], [28] is proposed to combine visual features and shape-based alignment using Kinect-style and normal RGB-D cameras, where visual and depth information are also combined for view-based loop-closure detection. Ramey *et al.* [29] integrate a low-cost commercial RGB-D sensor for gesture recognition, and extend the KinectFusion algorithm [31] by improving the RGB-D visual geometry algorithm. One main limitation of the RGB-D based works is that they are typically only suitable for indoor scenes. The proposed method, on the other hand, is able to extract object proposals in outdoor scenes.

Stereo-based Object Segmentation: Stereo-based object segmentation aims at segmenting objects from scenes using the stereo information. Bleyer *et al.* [32] introduce the concept of 3D scene-consistency into stereo matching by iteratively solving depth estimation and object extraction. This is a joint optimization problem, which iterats between two tasks so as to reduce the errors produced by individual tasks, with high computational complexity. This work is also related to RGBD-based works, as it extracts objects using only depth information. In [33], [34], an active vision system is presented for visual scene segmentation based on integration of several cues. These methods combine a set of foveal and peripheral cameras. Object hypotheses are generated through a stereo-based fixation process. Both [33] and [34] are based on measuring colors and binocular disparities, which inspire our work. However, stereo-based object segmentation focuses on directly segmenting the main object from the image, while object proposals aim at covering all generic objects in the image with a set of windows/segments.

III. Stereo Based Object Proposals

Similar to window scoring methods, the proposed stereo-based approach has two main parts: *proposal generation* and *objectness estimation*. Fig. 1 shows the framework of the proposed approach.

Proposal generation: The goal of proposal generation is to create a relatively small set of candidate bounding boxes that cover all objects in the image. To this end, we first generate initial bounding boxes using a stochastic strategy (Section III-B), which are then adaptively transformed to fit to the objects (Section III-C). Finally, a clustering method is applied to build a two-level hierarchy to separate the bounding boxes into the proposal and cluster levels (Section III-D).

Objectness estimation: Objectness is to describe the confidence that an image window contains an object of any category. In this work, objectness is not directly estimated for each bounding box as it is computationally expensive and bounding boxes with significant overlapping may produce unreliable results. We estimate objectness differently for the two levels of bounding boxes, for accuracy and efficiency (Section III-E).

Fig. 2. **Example of rough depth map enhancement**. (a) Color image. (b) Rough depth map. (c) and (d) Results after enhancement with different numbers of superpixels.

Two-level proposal ranking is applied to generate the final proposals (Section III-F).

A. Rough Depth Map Calculation

Given a stereo image pair, we can easily obtain a rough depth map using image correspondences. Most of the stereo matching algorithms may not be reliable in complex scenes, due to large depth range or flat regions. We use SIFT Flow [35] to generate the rough depth map because of its robustness in both indoor and outdoor scenes (Fig. 1(a)∼(b)).

Rough depth map enhancement: The estimated rough depth map typically has low quality especially for outdoor scenes, and the boundaries of objects may not be reliable. We enhance the depth map by utilizing the RGB information. Specifically, SLIC [36] is first used to over-segment the color image. As the color image and the rough depth map are aligned, we calculate the significant peaks of the depth histogram within each superpixel on the rough depth map. Suppose that a peak contains n_p pixels, the bins next to the peak contain n_l and n_r pixels respectively, and the superpixel contains N pixels. A peak is considered significant if the following conditions are satisfied:

$$n_p/N \geq \delta_1, \ \min\{n_p/n_l, n_p/n_r\} \geq \delta_2. \tag{1}$$

Finally, pixels inside the superpixel are assigned with the average depth value of the nearest peak (Fig. 2). This process can help smooth the rough depth map, remove tiny noisy areas, fill blank holes, and refine object boundary errors.

B. Initial Proposal Generation

To avoid generating hundreds of thousand of initial windows using the sliding window strategy as in [1], [2], [11], we only generate a relatively small set of N_b initial bounding boxes to reduce computational complexity using a stochastic strategy.

Suppose that the size of the image is $W \times H$. The upper left corner (x_1, y_1) of an initial proposal $\mathbf{b}$ is randomly generated in the range of $[1, W - T_b]$ for x_1 and $[1, H - T_b]$ for y_1, where T_b is a threshold. The width and height of $\mathbf{b}$ are also randomly generated in the range of $[T_b, W - x_1]$ and $[T_b, H - y_1]$, respectively, such that $W - x_1 \geq T_b$ and $H - y_1 \geq T_b$. Thus, the minimal size of $\mathbf{b}$ is $T_b \times T_b$ to avoid the bounding box being too small. This step produces the initial bounding box set as $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_{N_b}\}$ (Fig. 1(c)).

With the initial bounding boxes, we will transform them to adapt and fit to the objects (Section III-C). As demonstrated in our results, such a stochastic strategy achieves better results compared with the popular sliding window strategy since it ensures diversity (see Section V-B).

Initial proposal generation enhancement: With the RGB information, the initial proposal generation can be further improved. As suggested by [11], a patch with dense edges has higher probability to contain objects. Thus, we propose a method to leverage this observation. The image is divided into 8×8 regular patches and r_i is the summation of all edge values for those pixels within the ith patch, where the edge values are calculated using Structured edges [18]. The center of an initial bounding box is associated with a probability E_i of falling into the ith patch. E_i is defined as

$$\begin{aligned} E_i &= \frac{r_i}{\sum_{j=1}^{64} r_j}, \ i = 1, 2, ..., 64, \\ E_0 &= 0. \end{aligned} \tag{2}$$

Specifically, we generate a number ρ between 0 and 1 according to uniform distribution, and the proposal center is considered as falling into the kth patch if $\sum_{a=0}^{k-1} E_a < \rho \leq \sum_{b=0}^{k} E_b$. Then, the width and height of the proposal are computed using a similar idea. This way of distributing the initial bounding boxes conforms to the distribution of edge density, thus leads to a more accurate set of initial bounding boxes.

C. Adaptive Transformation

Adaptive transformation is proposed to convert the initial bounding boxes into better ones based on the depth map and the color image. Considering the fact that different parts of an object are likely on the same depth plane. In other words, a region with similar depth values possibly contains an object. Our goal is to find all these regions, and to minimize the influence from unreliable depth values.

Given an initial bounding box $\mathbf{b} \in \mathbf{B}$, we compute the histogram of depth values inside $\mathbf{b}$. The bin with the largest value is marked as $\mathbf{f}$, which represents the significant information of $\mathbf{b}$. Other bins are ignored and considered as noises. Then, we transform $\mathbf{b}$ by adaptively expanding or shrinking its boundaries such that $\mathbf{b}$ fits $\mathbf{f}$ better, as shown in Algorithm 1.

Efficient adaptive transformation: We utilize integral images to reduce the computational complexity, and the number of integral images is equal to the number of bins. For each bin $\mathbf{f}$, we first generate a binary image with the same size as the rough depth map, and 1 is assigned if the corresponding pixel falls into $\mathbf{f}$ and 0 otherwise. We then compute the integral image corresponding to $\mathbf{f}$ using the binary image. Hence, the number n can be easily computed with simple operations similar to [1].

Adaptive transformation enhancement: Color information can be used to enhance the adaptive transformation as the rough depth map is not reliable in some cases. For example, the depth values of a train moving towards the camera may vary significantly, and we may falsely divide the train into several proposals as it may be divided into several depth planes. Hence, we further perform adaptive transformation on the color image but with different thresholds. For each initial bounding box, we conduct RGB and depth based transformation individually. Thus, we obtain more transformed proposals, as a complement to the unreliable depth map (Fig. 3). Note that the depth and RGB information have their own advantages and disadvantages. It is difficult to decide which proposal is more suitable for objectness estimation. Hence, we retain both RGB-based and depth-based proposals. As the set is small, the extra computational cost is negligible.

Input: the depth map, the initial proposal set $\mathbf{B}$
Output: transformed proposal set $\mathbf{P}$
$\mathbf{P} = \varnothing$;
for *each* $\mathbf{b} \in \mathbf{B}$ **do**
  calculate the histogram of depth values inside $\mathbf{b}$;
  the bin with the largest value is marked as $\mathbf{f}$;
  for *each boundary of* $\mathbf{b}$ **do**
    count the number of pixels that fall into $\mathbf{f}$ as n;
    if $n < T_t$ **then**
      shrink the boundary towards proposal center and update n until $n \geq T_t$ is satisfied;
    end
    else if $n > T_t$ **then**
      expand the boundary and update n until $n \leq T_t$ is satisfied or reaches image boundary;
    end
  end
  $\mathbf{b} \rightarrow \mathbf{p}$ after transformation, and insert $\mathbf{p}$ into $\mathbf{P}$;
end
remove the repeated ones in $\mathbf{P}$; thus
$\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_{N_p}\}$, *s.t.* $N_p \leq N_b$ (see Fig. 1(d));

Algorithm 1: Pseudo-code for Adaptive Transformation.

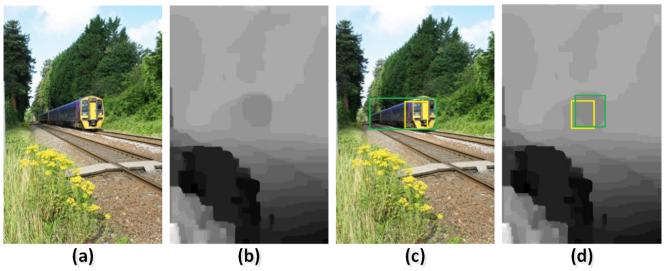


Fig. 3. **Example of an object with a large depth range but consistent color**. (a) Color image. (b) Rough depth map. (c) and (d) Results after transformation (yellow: initial bounding box, green: transformed bounding box).

D. Two-level Proposal Hierarchy

Although repeated proposals have been removed, some proposals may still be similar and have large overlapping ratios to each other when the total number of proposals, N_p, is large. This problem is mainly due to the adaptive transformation step, as different initial bounding boxes may result in similar proposals. On the other hand, we want to keep these similar proposals in order to achieve tight bounding boxes. Scoring all these proposals is computationally expensive, and non-

maximum suppression (NMS) is infeasible as objectness has not been estimated yet.

To solve this problem, a clustering method is adopted to construct a two-level hierarchy. At the proposal level, it stores the information of all proposals (Fig. 1(f)). At the cluster level, it stores the information of all cluster coordinates (Fig. 1(e)).

- ***Proposal level***: Each proposal $\mathbf{p}$ is represented by a 4D vector, i.e., the coordinates of the upper left and the bottom right corners.
- ***Cluster level***: We apply K-means clustering to divide all proposals based on their 4D vectors into clusters $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_{N_c}\}$, where N_c is the total number of clusters and each cluster $\mathbf{c}_i = \{\mathbf{p}_1^i, \mathbf{p}_2^i, ..., \mathbf{p}_{n_i}^i\}$, $s.t.\ n_1 + n_2 + ... + n_i = N_p,\ i = 1, 2, ...N_c$, and $N_c \ll N_p$. For each cluster $\mathbf{c}_i$, we compute the cluster coordinate $\mathbf{o}_i$, which is also a 4D vector, as the clustered proposal.

There are two advantages of this two-level hierarchy. First, abstract information extracted from $\mathbf{P}$ can be retained, as each $\mathbf{o}_i$ represents many similar proposals, and redundancy can be avoided with a smaller set $\mathbf{C}$ in the cluster level. Second, objectness estimation can be accelerated, as the scores for those similar proposals can be easily estimated, while the diversity can be retained in the proposal level (see Section III-E).

E. Objectness Estimation

Three novel stereo-based cues "exactness", "focus" and "distribution" are proposed to estimate objectness. Note that cluster coordinates $\mathbf{o}_i$ of $\mathbf{c}_i$ are 4D vectors and can be regarded as special proposals. Thus, these cues can be applied to both proposal and cluster levels. For simplicity the term "proposal" represents both in this subsection.

The depth map calculated from the stereo image pair is first normalized to 0∼1 and divided into several depth planes according to the depth values. For each proposal, its corresponding depth plane $\mathbf{f}$ is defined as the plane that most pixels belong to. Hence, a proposal can only be associated with a specific $\mathbf{f}$. The three cues are defined as follows.

- ***Exactness***: Our observation is that if a proposal exactly consists of an object, the surrounding region $\mathbf{T}$ should contain little portion of the object, where $\mathbf{T}$ is defined as the enlargement of the proposal by a factor of θ ($\theta > 1$) in all directions. Let the total number of pixels covered by $\mathbf{T}$ be $a_{\mathbf{r}}$. We count the number of pixels $n_{\mathbf{r}}$ in $\mathbf{T}$ belonging to depth plane $\mathbf{f}$. Preferably, this number should be small. "Exactness" is then computed as:

$$S_e = \frac{a_{\mathbf{r}} - n_{\mathbf{r}}}{(b_w \times b_h)^{\kappa}}, \tag{3}$$

  where b_w and b_h are the width and height of the bounding box. Note that noise can appear in the depth estimation step (i.e., the depth values within a small region are falsely estimated), and this noise typically appears in small regions. Proposals with larger areas are assigned high priority to confidently estimate the objectness by setting $0 < \kappa < 1$. Larger S_e indicates that $\mathbf{T}$ contains a smaller portion of the object, and hence the proposal has a higher possibility that it exactly consists of an object.

- ***Focus***: Photographers tend to arrange salient objects in a different depth level from the background. This means that regions with small depth values more likely belong to an object. Let $\mathbf{M}$ be the set of all pixels that fall within $\mathbf{f}$. As the largest depth value after normalization is 1, representing the background, we formulate "Focus" to describe the distance of each pixel in $\mathbf{M}$ from the background as:

$$S_f = \frac{\sum_{\mathbf{m}\in\mathbf{M}}(1 - d_{\mathbf{m}})}{(b_w \times b_h)^{\kappa}}, \tag{4}$$

  where $d_{\mathbf{m}}$ is the depth value of pixel $\mathbf{m}$. Larger S_f indicates a higher probability that the proposal is close to the camera. Note that most objects are not rectangular. Thus, we ignore regions within the proposal that fall outside $\mathbf{f}$. Again, large proposals are assigned high priority in this cue.

- ***Distribution***: Both [1] and [11] suggest that edges play an important role in objectness estimation, and edges near the proposal boundary are more important than those near the center. They use a binary weight (0 or 1) in the close-to-boundary region. However, this region is manually defined and the transition between 0 and 1 is not smooth. In this work, we utilize a weight mask, which has the same size as the proposal, to indicate the pixel importance based on the relative positions of the pixels in the proposal. The weight w_{ij} of a pixel at location (i, j) decreases with distance from proposal boundaries to the center as follows:

$$w_{ij} = \begin{cases} \frac{\lfloor \max\{|j-c_y|, |i-c_x|-\upsilon\} \rfloor}{\lceil \min\{b_w, b_h\}/2 \ - \ 1 \rceil} & b_w \geq b_h \\ \frac{\lfloor \max\{|j-c_y|-\upsilon, |i-c_x|\} \rfloor}{\lceil \min\{b_w, b_h\}/2 \ - \ 1 \rceil} & b_w < b_h \end{cases}, \tag{5}$$

  where $\upsilon = |b_w - b_h|/2$, $(c_x, c_y) = (b_w/2, b_h/2)$ is the relative position of the proposal center. "Distribution" is then formulated as:

$$S_d = \sum_{i=1}^{b_h} \sum_{j=1}^{b_w} w_{ij} \times e_{ij}, \tag{6}$$

  where e_{ij} is the edge response. Larger S_d indicates a higher probability that the proposal consists of an object as edges tend to appear near the proposal boundary. e_{ij} is mainly derived from structured edges [18] by further considering occlusions in the stereo image pair, i.e., some edges of an object may appear in one image but not in the other. The occluded edges are omitted as they may not be reliable. To detect occlusions, we find the edge group correspondences between two images. The edge groups are first generated in the way similar to [11] for both images. The correspondences and confidence values are computed by considering the data, small displacement and smoothness terms similar to SIFT Flow [35], except the spatial pyramid strategy as we only search the adjacent areas for efficiency. Then, the edges with low confidence values are considered as occlusions and the edge responses are set to $e_{ij} = 0$; otherwise e_{ij} remains the same as the edge response from structured edges [18].

Input: $\mathbf{C}'$. The objectness of $\mathbf{c}'_i$ is larger than that of $\mathbf{c}'_j$, for $i < j$. In each cluster $\mathbf{c}'_k$, the objectness of $\mathbf{p}'^k_m$ is larger than that of $\mathbf{p}'^k_n$, for $m < n$.
Output: ranked proposals list $\mathbf{R}$.
Step 1: $\mathbf{R} = \varnothing$.
Step 2: The first ranked proposal from each cluster is added to $\mathbf{R}$, thus $\mathbf{R} = \{\mathbf{p}'^1_1, \mathbf{p}'^2_1, ..., \mathbf{p}'^{N'_c}_1\}$. The added proposals are removed from the original clusters.
Step 3: Keep moving the new first ranked proposal from each non-empty cluster to $\mathbf{R}$.
Step 4: The process is finished when all clusters are empty or $\mathbf{R}$ contains enough proposals.

Algorithm 2: Pseudo-code for Proposal Ranking.

Min-max normalization is performed on S_e, S_f and S_d to ensure that they range between 0 and 1. In our two-level hierarchy, the cluster level is more important as it decides if the regions contain objects, while the proposal level mainly affects the proposal tightness. For efficiency, we compute objectness $S_{\mathbf{c}}$ with all cues in the cluster level $\mathbf{C}$, and objectness $S_{\mathbf{p}}$ with "exactness" only in the proposal level $\mathbf{P}$ as:

$$\begin{aligned} S_{\mathbf{c}} &= \alpha_e S_e + \alpha_f S_f + (1 - \alpha_e - \alpha_f) S_d \\ S_{\mathbf{p}} &= S_e, \end{aligned} \tag{7}$$

where α_e and α_f are the weight coefficients to indicate the importance of the three cues. Since "exactness" can best distinguish similar proposals as it is sensitive to small changes, we use it to estimate objectness in the proposal level.

We then apply NMS in the cluster level to remove any clustered proposal if it overlaps (intersection-over-union (IoU) larger than 0.75) with another clustered proposal with a higher objectness score. If a clustered proposal is removed, all the proposals belonging to the cluster are also removed. The remaining clustered proposals and proposals are sorted and saved in descending order of objectness, i.e., $\mathbf{C}' = \{\mathbf{c}'_1, \mathbf{c}'_2, ..., \mathbf{c}'_{N'_c}\}$, where N'_c is the number of remaining clusters and $N'_c \leq N_c$, with each remaining cluster $\mathbf{c}'_i = \{\mathbf{p}'^i_1, \mathbf{p}'^i_2, ..., \mathbf{p}'^i_{n'_i}\}$, $i = 1, 2, ..., N'_c$.

F. Proposal Ranking

The goal of proposal ranking is to obtain good proposals from the two-level hierarchy objectness scores, where top ranked proposals are supposed to have higher probabilities of containing objects. Instead of comparing object proposals across different clusters, we compare them using a two-level approach. The main idea is that we should prevent proposals from the same cluster having similar rankings in the final list to ensure diversity. The rankings of dissimilar proposals from different clusters are based on objectness of the corresponding cluster coordinates. Algorithm 2 shows the pseudo code for proposal ranking.

This method can enhance the diversity of top-ranking proposals in $\mathbf{R}$ and improve the ranking performance, as it prevent proposals from the same cluster having similar rankings and maintains the importance of clusters at the same time.

IV. Integration with RGB-based Methods

The proposed approach can be easily integrated with existing RGB-based methods. Two integration approaches are proposed here, depending how stereo and RGB proposals are integrated: *Late Integration* – two types of proposals are directly integrated, and *Early Integration* – stereo proposals are integrated with RGB-based objectness function.

For simplicity, we define $\mathbf{P}_{rgb}$ as the set of all proposals generated by the RGB-based algorithms, $\mathbf{F}_{RGB}(\cdot)$ as the function to estimate RGB-based objectness, and $S_{RGB \leftarrow rgb}$ as the corresponding objectness value of $\mathbf{p} \in \mathbf{P}_{rgb}$. For the proposed stereo-based approach, $\mathbf{P}_d$, $\mathbf{F}_D(\cdot)$ and $S_{D \leftarrow d}$ are defined similarly and are already known. Our goal here is to integrate $\mathbf{P}_{rgb}$ and $\mathbf{P}_d$ to generate new proposals.

Late Integration: If only the executables of the existing methods are available, $\mathbf{F}_{RGB}(\cdot)$ will not be available although $\mathbf{P}_{rgb}$ and $S_{RGB \leftarrow rgb}$ can be obtained. $\mathbf{F}_D(\cdot)$ is used to calculate depth-based objectness for $\mathbf{P}_{rgb}$ as:

$$S_{D \leftarrow rgb} = \mathbf{F}_D(\mathbf{p} \in \mathbf{P}_{rgb}). \tag{8}$$

For each $\mathbf{p} \in \mathbf{P}_d$, we compare it with all proposals in $\mathbf{P}_{rgb}$ to select one $\bar{\mathbf{p}}_{rgb}$ having maximum IoU λ with $\mathbf{p}$. We then estimate $S_{RGB \leftarrow d}$ as:

$$S_{RGB \leftarrow d} = \bar{S}_{RGB} \times \lambda, \tag{9}$$

where $\bar{S}_{RGB}$ is the RGB-based objectness of $\bar{\mathbf{p}}_{rgb}$. Integrated objectness S_I is computed for $\mathbf{p} \in (\mathbf{P}_{rgb} \cup \mathbf{P}_d)$:

$$S_I = \begin{cases} {S_{D \leftarrow rgb}}^{\omega_1} \times {S_{RGB \leftarrow rgb}}^{1-\omega_1} & \mathbf{p} \in \mathbf{P}_{rgb} \\ {S_{D \leftarrow d}}^{\omega_1} \times {S_{RGB \leftarrow d}}^{1-\omega_1} & \mathbf{p} \in \mathbf{P}_d \end{cases}, \tag{10}$$

where ω_1 is a weight coefficient. The advantage of the above method is that it does not require $\mathbf{F}_{RGB}(\cdot)$ to be available.

Early Integration: If the source codes of the existing methods are available, $\mathbf{P}_{rgb}$, $S_{RGB \leftarrow rgb}$ and $\mathbf{F}_{RGB}(\cdot)$ will all be available. $S_{D \leftarrow rgb}$ is computed with Eq. 8, and $S_{RGB \leftarrow d}$ is directly computed from $\mathbf{F}_{RGB}(\cdot)$ as:

$$S_{RGB \leftarrow d} = \mathbf{F}_{RGB}(\mathbf{p} \in \mathbf{P}_d). \tag{11}$$

Integrated objectness S_I is computed for $\mathbf{p} \in (\mathbf{P}_{rgb} \cup \mathbf{P}_d)$:

$$S_I = \begin{cases} {S_{D \leftarrow rgb}}^{\omega_2} \times {S_{RGB \leftarrow rgb}}^{1-\omega_2} & \mathbf{p} \in \mathbf{P}_{rgb} \\ {S_{D \leftarrow d}}^{\omega_2} \times {S_{RGB \leftarrow d}}^{1-\omega_2} & \mathbf{p} \in \mathbf{P}_d \end{cases}, \tag{12}$$

where ω_2 is a weight coefficient. This method requires to have access to the $\mathbf{F}_{RGB}(\cdot)$ code. The advantage is that the computed $S_{RGB \leftarrow d}$ is more reliable.

Finally, we sort all RGB and depth proposals based on S_I and apply NMS to remove similar ones, to produce the final ranked proposals.

V. Experiments

To evaluate the proposed method, we set $T_b = 20$ such that the minimal size of initial bounding boxes is 20×20, and randomly generate $N_b = 10000$ initial bounding boxes for each depth map. K-means is used to generate $N_c = 200$ clusters for the two-level hierarchy. We compute the histogram with 20 bins, which represent 20 depth planes in this work. α_e and

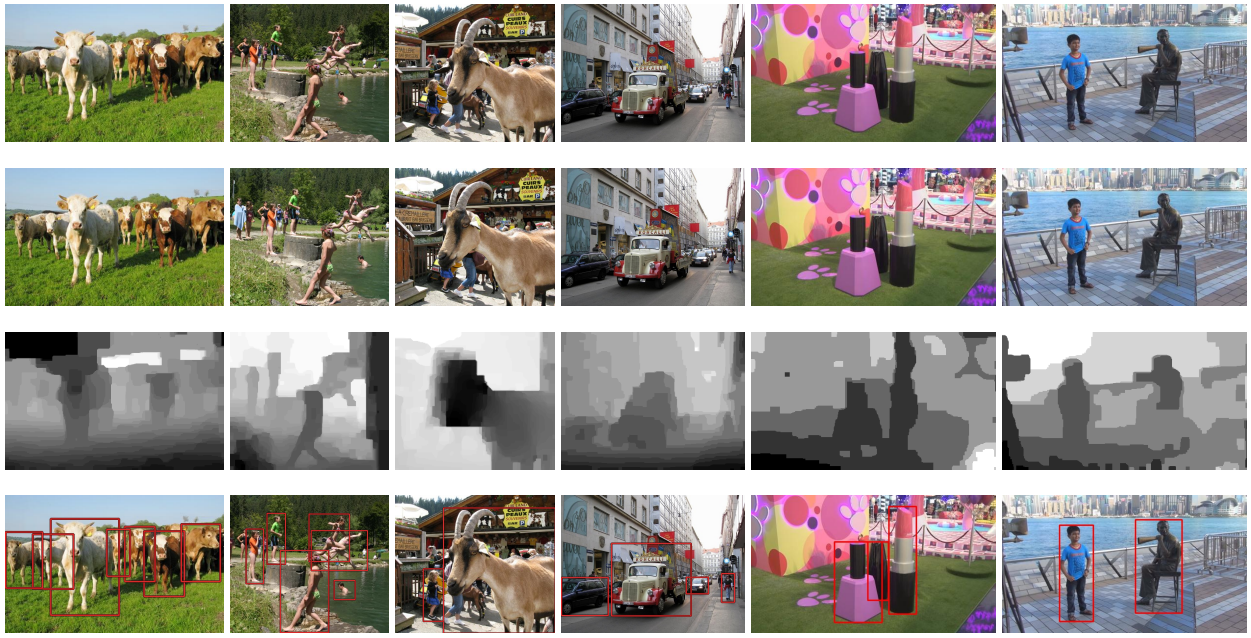


Fig. 4. **Examples of the proposed StereoObj dataset.** From top to bottom: left-view images, right-view images, rough depth maps, ground truth (red).

α_f in Eq. 7 are heuristically set to 0.4 and 0.3, respectively, to slightly bias to "exactness". We conduct all experiments on a PC with an i7 3GHz CPU and 18GB RAM. Our work is implemented using Matlab. Detection rate vs. number of proposals at different IoU thresholds (0.5, 0.6, 0.7, 0.9) is used as the main evaluation metric for object proposal detection performance. We have compared the proposed method with the state-of-the-arts [1], [2], [6], [11], [37], [38] against multiple factors.

For the rest of this section, we first introduce the proposed stereo dataset in Section V-A. We then examine different initial sampling strategies in Section V-B and the effects of different components in Section V-C. In Sections V-D, V-E and V-F, we evaluate the proposed method on the constructed StereoObj dataset, the Cityscapes dataset [39], and the KITTI dataset [40], respectively. Finally, we evaluate the performance improvement of integrating the proposed method to a general color-based framework in Section V-G.

A. Datasets

As the PASCAL VOC dataset [41] is for single color images, it is difficult to be applied to our work. Although the Middlebury Stereo Datasets [42] provides high-resolution stereo image pairs with subpixel-accurate ground truth, it contains only 33 stereo image pairs: 10 for training, 10 for testing, plus additional 13. The number of images is too small and they are all taken in indoor scenes with a limited depth range, which are not practical enough for object detection in real world scenes.

Hence, we build a stereo dataset, called StereoObj dataset, with both indoor and outdoor scenes to evaluate the performance and to facilitate further object-related applications. The dataset comprises of 400 stereo image pairs, which are taken from different scenes including beaches, cities, streets, buildings, sculptures, sports, rooms and malls. These images contain different challenging factors, e.g., objects occlusion, persons standing shoulder to shoulder, cluttered background, tiny and distant objects, and objects with large depth variation. The locations of objects are labeled manually, and each image contains at least one object. Fig. 4 shows some examples of the StereoObj dataset.

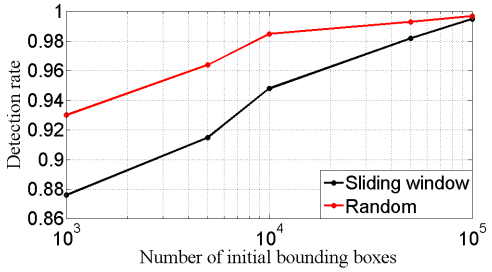


Fig. 5. **Effects of different initial sampling strategies**. For a given number of initial bounding boxes, we evaluate the detection rate for the top-1000 ranked proposals when IoU = 0.5. "Sliding window" uses the sliding window strategy to generate the initial bounding boxes. "Random" uses the method proposed in Section III-B to generate the initial bounding boxes.

In addition to StereoObj, we further evaluate the proposed method on the latest Cityscapes [39] and the challenging KITTI [40] datasets. The Cityscapes dataset contains two categories in instance-level: vehicles and humans; we use detection rate vs. number of proposals as the evaluation metric. The KITTI dataset contains three object classes: car, pedestrian and cyclist; we follow the evaluation setting of 3DOP [43] for fair comparison.

B. Examining Different Initial Sampling Strategies

As mentioned in Section III-B, we propose a stochastic strategy to generate a relatively small number of initial bounding boxes to reduce computational complexity. This

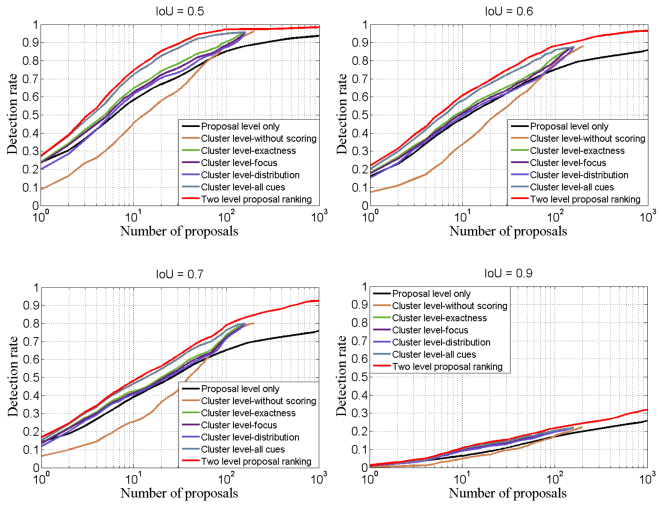


Fig. 6. **Effects of proposed cues on the StereoObj dataset**. "Proposal level only" shows the results when the objectness of proposals is directly estimated without clustering. "Cluster level-without scoring" shows the results when the clustered proposals are randomly ranked without scoring. "Cluster level-exactness", "Cluster level-focus" and "Cluster level-distribution" show the results when the corresponding cues are used in objectness estimation at the clustered proposals. "Cluster level-all cues" shows the results when all three cues are used in objectness estimation. Note that all these variants do not include proposal level ranking. "Two level proposal ranking" shows the results when using all the proposals from our two-level hierarchy, with both cluster and proposal level ranking.

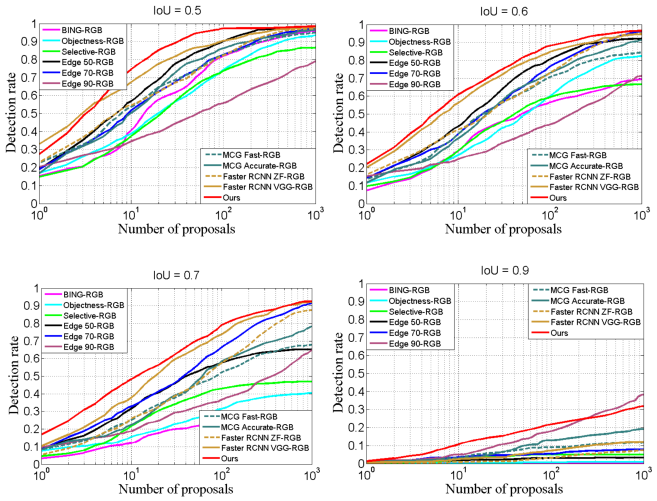


Fig. 7. **Comparison with RGB-based methods on the StereoObj dataset: BING [2], Objectness [1], Selective Search [6], Edge Boxes [11], MCG [37], and Faster R-CNN [38]**. Variants of Edge Boxes ($\delta = 0.5, 0.7$ and 0.9, denoted as Edge 50, Edge 70 and Edge 90) are tested. Two versions of MCG, the fast version (MCG Fast) and original version (MCG Accurate) are evaluated. Two models of Faster R-CNN (ZF [44] and VGG [45]) are also evaluated. We feed the left images to the existing RGB-based methods.

avoids initializing with exhaustive sliding window searching [1], [2], [11]. In Fig. 5, we show the improvement brought by random sampling. As can be seen, the randomly generated initial bounding boxes after adaptive transformation are able to achieve better performance, when the number of initial bounding boxes is small. In other words, most of the bounding boxes generated by sliding windows are not necessary.

C. Analyzing the Effects of Different Components

Fig. 6 compares the performances of different cues and key steps for proposals generation and objectness estimation, on the StereoObj dataset. "Cluster level-exactness", "Cluster level-focus" and "Cluster level-distribution" denote the performances when applying the corresponding cues to the clustered proposals. As can be seen, each proposed cue helps improve objectness estimation as all three curves show significant improvements over random ranking (i.e., "Cluster level-without scoring"). "Cluster level-all cues" dentoes the performance of combining all three cues. It outperforms all three individual ones. The proposed two-level hierarchy is also examined (i.e., "Two level proposal ranking"). As clustered proposals only capture the abstract locations of the objects, they typically not fit the objects well. On the other hand, the proposals belonging to the same cluster in the proposal level are not diverse enough. By combining the proposals from the two levels, the proposed objectness estimation and ranking schemes are able to select the tightest proposals that cover the objects.

D. Comparing on the StereoObj Dataset

We compare the proposed method with the state-of-the-art RGB-based methods [1], [2], [6], [11], [37], [38] on the

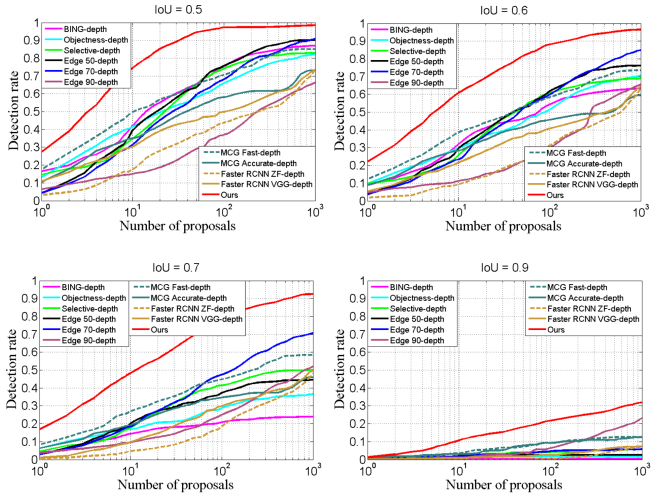
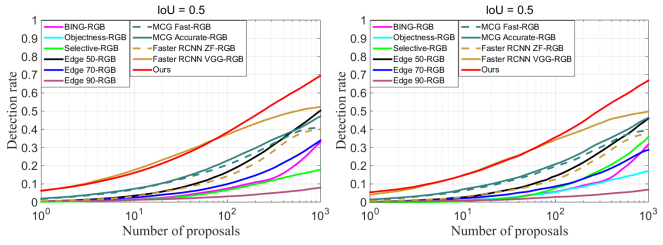


Fig. 8. **Comparison with RGB-based methods on the StereoObj dataset when depth maps are used as input: BING [2], Objectness [1], Selective Search [6], Edge Boxes [11], MCG [37], and Faster R-CNN [38]**. The variants of Edge Boxes ($\delta = 0.5, 0.7$ and 0.9, denoted as Edge 50, Edge 70 and Edge 90) are tested.

StereoObj dataset. For the existing RGB-based methods, we use the left image as input. (Our experiment shows that the differences are slight when feeding the left or the right images to these methods in generating the proposals.) Fig. 7 shows that the performance of the proposed method is better than the color-based methods in all IoU settings, including Faster R-CNN [38].

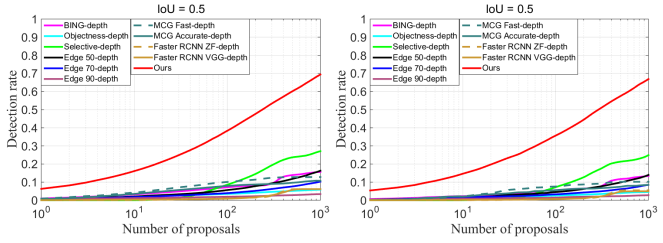
We have also compared the proposed method with these existing methods when they are fed with the rough depth maps, instead of color images. Fig. 8 shows that these methods with depth maps as input perform worse than those with color images. The main reason is that the estimated depth maps are inaccurate and unreliable. Hence, the generated proposals

Fig. 9. **Qualitative results of the proposed method on the StereoObj dataset**. The ground truth bounding boxes are in red color. The corresponding best proposals are in green color, obtained from the top N proposals. From top to bottom: $N = 1$, $N = 10$, $N = 100$ and $N = 1000$.



(a) Training set (b) Validation set

Fig. 10. **Comparison with RGB-based methods on the Cityscapes dataset: BING [2], Objectness [1], Selective Search [6], Edge Boxes [11], MCG [37], and Faster R-CNN [38]**. We feed the left images to the existing RGB-based methods.



(a) Training set (b) Validation set

Fig. 11. **Comparison with RGB-based methods on the Cityscapes dataset when depth maps are used as input: BING [2], Objectness [1], Selective Search [6], Edge Boxes [11], MCG [37], and Faster R-CNN [38]**.

are not capable of covering all objects in the scene. The performance drop of Faster R-CNN indicates that the features learned from color images are not suitable for depth maps.

Fig. 9 shows some qualitative results of the proposed method on the StereoObj dataset, with respect to different numbers of proposals. We can see that the performance is still good even with a small number of proposals.

E. Comparing on the Cityscapes Dataset

Cordts *et al.* [39] establish the Cityscapes dataset, which contributes to semantic understanding of urban street scenes. Although they include the pixel-level and instance-level semantic segmentation tasks, these tasks differ from our objective and thus the proposed metric cannot be directly used for evaluation. Hence, we apply the same metric (detection rate vs. number of proposals) on the Cityscapes dataset, by segmenting each instance using a ground truth (GT) bounding box. As the vehicles and humans are most important for autonomous driving, Cordts *et al.* [39] only identify the two categories in the instance-level. We have conducted comparisons on the training and validation sets, which contain 2975 and 500 fine-annotated images, respectively, since the annotations for the test set are unavailable and their tasks are not suitable for object proposals.

Fig. 10 and 11 show the performances when the left images and the rough depth maps, respectively, are used as input for the RGB-based methods. IoU threshold is chosen 0.5 to determine whether an instance is detected. The results demonstrate that the proposed method achieves superior performance compared with the RGB-based methods on both training and validation sets. Same as the conclusion drawn for the StereoObj dataset, these methods with depth maps as input perform worse than those with color images. Fig. 12 shows some examples for qualitative evaluation.

F. Comparing on the KITTI Dataset

To evaluate the proposed method on the KITTI dataset [40], we follow the evaluation steps in 3DOP [43] to use the moderate setting for the object classes of cars, pedestrians and cyclists. For each ground-truth (GT) object, we regard

Fig. 12. **Qualitative results of the proposed method on the Cityscapes dataset**. The first row shows the GT bounding boxes in blue. The other rows show the results of different methods when 1000 top ranked proposals are used. From top to bottom: BING [2], Objectness [1], Selective Search [6], Edge Boxes (Edge 50) [11], MCG (Accurate) [37], Faster R-CNN (VGG) [38] and our method. Red rectangles indicate corresponding instances that are successfully detected (i.e., IoU > 0.5), while green rectangles indicate instances that fail to be detected by any proposals. The numbers on top of each rectangle shows the IoU overlap with the instance.

the proposal with the highest IoU value as its proposal and the recall is a success if this IoU value exceeds 70% for cars, and 50% for pedestrians and cyclists. The proposed method is compared with the following baselines whose results were reported in 3DOP: BING [2], Selective Search [6], Edge Boxes [11], MCG [37] and 3DOP [43]. Faster R-CNN [38] is also evaluated, and the VGG model [45] is used to generate the generic proposals with the default parameters. Fig. 13 shows recall as a function of the number of proposals. The results demonstrate that the proposed method achieves superior performance compared with RGB-based methods. 3DOP is the only one consistently outperforms our method. The main reason is that 3DOP has a different focus from ours (3D vs. stereo), and it is designed to capture 3D object proposals. Another reason is that it is a deep learning approach trained with additional data.

G. Integrating with RGB-based Methods

We evaluate the integration performance proposed in Section IV when IoU threshold is chosen 0.5 and 0.7 respectively on the StereoObj dataset. As shown in Fig. 14 and 15, the

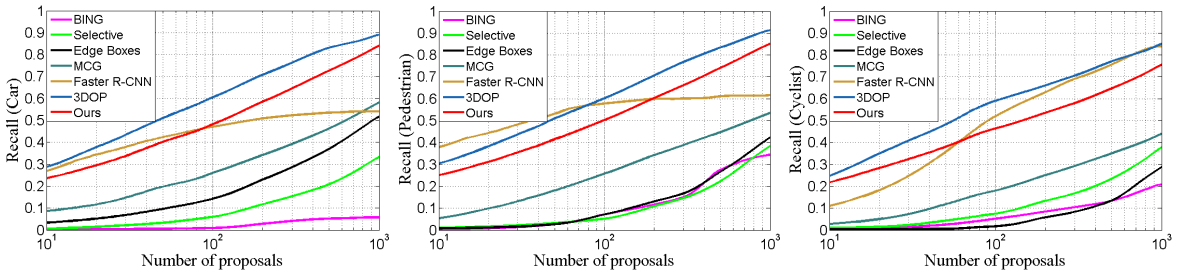


Fig. 13. **Comparison on the KITTI dataset [40] using the moderate setting: BING [2], Selective Search [6], Edge Boxes [11], MCG [37], 3DOP [43] and Faster R-CNN [38].** Left: car. Middle: pedestrian. Right: cyclist.

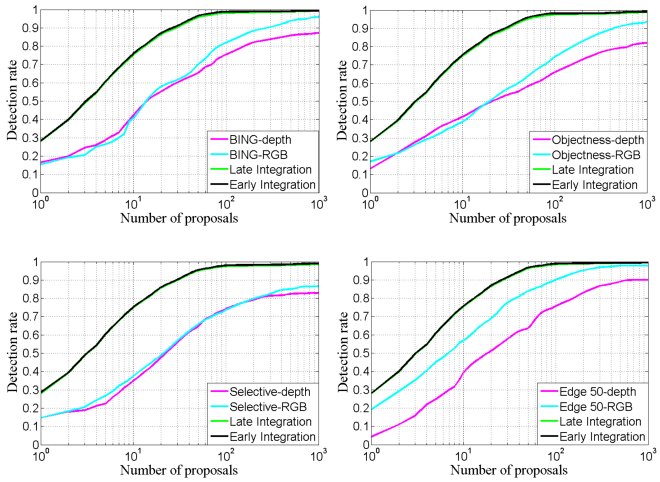


Fig. 14. **Evaluation on the two integration approaches with existing RGB-based methods under IoU threshold=0.5: BING [2], Objectness [1], Selective Search [6], and Edge Boxes [11].** Each diagram shows results of one existing method, using depth maps (i.e., "XXXX-depth") or color images (i.e., "XXXX-RGB") as input, and the two integration approaches "Late Integration" and "Early Integration".

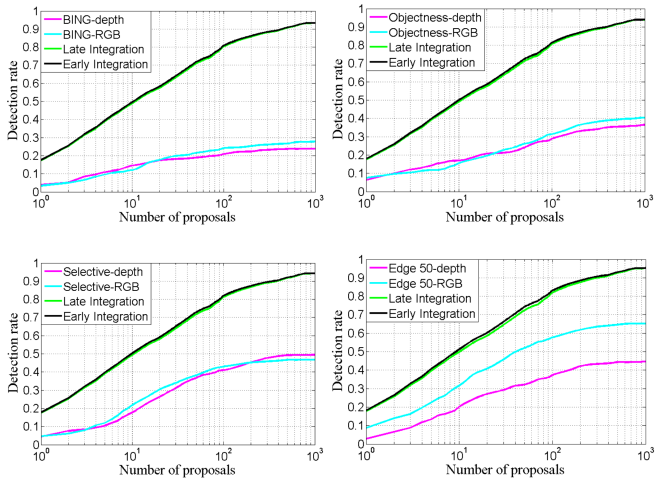


Fig. 15. **Evaluation on the two integration approaches with existing RGB-based methods under IoU threshold=0.7.**

improvements brought by the two types of integration are significant. Hence, the proposed stereo framework can be considered as a complement to the existing methods. We also note that "Early Integration" always performs slightly better than "Late Integration", as $S_{RGB \leftarrow d}$ is more reliable.

VI. Conclusion

In this paper, we leverage stereopsis for generating object proposals to combat cluttered background. We propose a framework that involves adaptive transformation, two-level hierarchy construction, objectness estimation and proposal ranking, to generate ranked proposals. We have also constructed the StereoObj dataset with 400 labeled stereo image pairs containing both indoor and outdoor scenes for evaluation. Experiments show that the proposed method performs significantly better than existing RGB-based methods. Two integration strategies with existing RGB-based methods are discussed, and both obtain superior performance over the original RGB-based methods. As a future work, we are currently investigating the application of stereopsis with the proposed cues in other tasks.

Acknowledgments

We thank Flickr user "3D Shoot" [1] for the permission to use his photos to construct our dataset. The work was supported in part by the National Natural Science Foundation of China under Grant No. 61271434, No. 61232013, and by Beijing Advanced Innovation Center for Imaging Technology under Grant No. BAICIT-2016009. The work was also partially supported by an SRG from CityU (Ref. 7004416).

References

[1] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE TPAMI*, vol. 34, no. 11, pp. 2189–2202, 2012.
[2] M. Cheng, Z. Zhang, W. Lin, and P. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," in *CVPR*, 2014, pp. 3286–3293.
[3] I. Endres and D. Hoiem, "Category-independent object proposals with diverse ranking," *IEEE TPAMI*, vol. 36, no. 2, pp. 222–234, 2014.
[4] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.
[5] S. Manen, M. Guillaumin, and L. V. Gool, "Prime object proposals with randomized prim's algorithm," in *ICCV*, 2013, pp. 2536–2543.
[6] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013.
[7] H. Harzallah, F. Jurie, and C. Schmid, "Combining efficient object localization and image classification," in *ICCV*, 2009, pp. 237–244.
[8] C. Lampert, M. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," in *CVPR*, 2008, pp. 1–8.

[1] https://www.flickr.com/photos/62273460@N05/

[9] P. Siva and T. Xiang, "Weakly supervised object detector learning with model drift detection," in *ICCV*, 2011, pp. 343–350.

[10] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *ICCV*, 2009, pp. 606–613.

[11] C. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*, 2014, pp. 391–405.

[12] D. Hoiem, A. Efros, and M. Hebert, "Geometric context from a single image," in *ICCV*, 2005, pp. 654–661.

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

[14] B. Russell, W. Freeman, A. Efros, J. Sivic, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *CVPR*, vol. 2, 2006, pp. 1605–1614.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[16] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," in *ICCV*, 2013, pp. 17–24.

[17] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, "What makes for effective detection proposals?" *IEEE TPAMI*, vol. 38, no. 4, pp. 814–830, 2016.

[18] P. Dollár and C. Zitnick, "Structured forests for fast edge detection," in *ICCV*, 2013, pp. 1841–1848.

[19] J. Carreira and C. Sminchisescu, "CPMC: Automatic object segmentation using constrained parametric min-cuts," *IEEE TPAMI*, vol. 34, no. 7, pp. 1312–1328, 2012.

[20] I. Endres and D. Hoiem, "Category independent object proposals," in *ECCV*, 2010, pp. 575–588.

[21] C. Long, X. Wang, G. Hua, M. Yang, and Y. Lin, "Accurate object detection with location relaxation and regionlets re-localization," in *ACCV*, 2014, pp. 260–275.

[22] S. He and R. Lau, "Oriented object proposals," in *ICCV*, 2015, pp. 280–288.

[23] X. Chen, H. Ma, X. Wang, and Z. Zhao, "Improving object proposals with multi-thresholding straddling expansion," in *CVPR*, 2015, pp. 2587–2595.

[24] M. Bertamini, "Sensitivity to reflection and translation is modulated by objectness," *Perception*, vol. 39, no. 1, p. 27, 2010.

[25] M. Bleyer, C. Rother, P. Kohli, D. Scharstein, and S. Sinha, "Object stereo-joint stereo matching and object segmentation," in *CVPR*, 2011, pp. 3081–3088.

[26] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *ECCV*, 2014, pp. 345–360.

[27] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "RGB-D mapping: Using kinect-style depth cameras for dense 3D modeling of indoor environments," *IJRR*, vol. 31, no. 5, pp. 647–663, 2012.

[28] ——, "RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments," in *Experimental Robotics*, 2014, pp. 477–491.

[29] A. Ramey, V. González-Pacheco, and M. A. Salichs, "Integration of a low-cost RGB-D sensor in a social robot for gesture recognition," in *HRI*, 2011, pp. 229–230.

[30] T. Whelan, H. Johannsson, M. Kaess, J. Leonard, and J. McDonald, "Robust real-time visual odometry for dense RGB-D mapping," in *ICRA*, 2013, pp. 5724–5731.

[31] R. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *ISMAR*, 2011, pp. 127–136.

[32] M. Bleyer, C. Rhemann, and C. Rother, "Extracting 3D scene-consistent object proposals and depth from stereo images," in *ECCV*, 2012, pp. 467–481.

[33] M. Björkman and D. Kragic, "Active 3D scene segmentation and detection of unknown objects," in *ICRA*, 2010, pp. 3114–3120.

[34] ——, "Active 3D segmentation through fixation of previously unseen objects," in *BMVC*, 2010, pp. 119.1–119.11.

[35] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. Freeman, "SIFT Flow: Dense correspondence across different scenes," in *ECCV*, 2008, pp. 28–42.

[36] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE TPAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

[37] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *CVPR*, 2014, pp. 328–335.

[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.

[39] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.

[40] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, 2012, pp. 3354–3361.

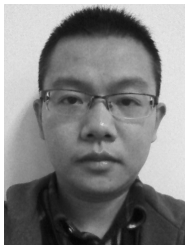
[41] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

[42] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Pattern Recognition*, 2014, pp. 31–42.

[43] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3D object proposals for accurate object class detection," in *NIPS*, 2015, pp. 424–432.

[44] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014, pp. 818–833.

[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.



Shao Huang is a joint Ph.D. student of University of Chinese Academy of Sciences, CAS, China, and City University of Hong Kong, Hong Kong. He obtained his B.E. degree from Nankai University in 2011.

His research interests include multimedia technology, pattern recognition, image processing, computer vision, and deep learning.

Weiqiang Wang received the B.E. and M.E. degrees in computer science from Harbin Engineering University, Harbin, China, in 1995 and 1998, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2001.

He is currently a professor with School of Computer and Controlling Engineering at University of Chinese Academy of Sciences, and a guest researcher with Beijing Advanced Innovation Center for Imaging Technology, Capital Normal University. His research interests include multimedia content analysis, computer vision and machine learning.

Shengfeng He is an Associate Professor in the School of Computer Science and Engineering at South China University of Technology. He obtained his B.Sc. degree and M.Sc. degree from Macau University of Science and Technology in 2009 and 2011 respectively, and the Ph.D degree from City University of Hong Kong in 2015. His research interests include computer vision, image processing, computer graphics, and deep learning.

Rynson W.H. Lau received his Ph.D. degree from University of Cambridge. He was on the faculty of Durham University and is now with City University of Hong Kong.

Rynson serves on the Editorial Board of Computer Animation and Virtual Worlds. He has served as the Guest Editor of a number of journal special issues, including ACM Trans. on Internet Technology, IEEE Trans. on Multimedia, IEEE Trans. on Visualization and Computer Graphics, and IEEE Computer Graphics & Applications. He has also served in the committee of a number of conferences, including Program Co-chair of ACM VRST 2004, ACM MTDL 2009, IEEE U-Media 2010, and Conference Co-chair of CASA 2005, ACM VRST 2005, ACM MDI 2009, ACM VRST 2014. Rynson's research interests include computer graphics and computer vision.