SEMANTIC CONTEXT AND DEPTH-AWARE OBJECT PROPOSAL GENERATION

Haoyang Zhang[⋆,†], *Xuming He*[⋆,†], *Fatih Porikli*[⋆,†], *Laurent Kneip*[†]

[⋆]NICTA, Canberra; [†]Australian National University, Canberra

ABSTRACT

This paper presents a context-aware object proposal generation method for stereo images. Unlike existing methods which mostly rely on image-based or depth features to generate object candidates, we propose to incorporate additional geometric and high-level semantic context information into the proposal generation. Our method starts from an initial object proposal set, and encode objectness for each proposal using three types of features , including a CNN feature, a geometric feature computed from dense depth map, and a semantic context feature from pixel-wise scene labeling. We then train an efficient random forest classifier to re-rank the initial proposals and a set of linear regressors to fine-tune the location of each proposal. Experiments on the KITTI dataset show our approach significantly improves the quality of the initial proposals and achieves the state-of-the-art performance using only a fraction of original object candidates.

Index Terms— Object proposal, object detection, scene context, 3D scene.

1. INTRODUCTION

Generating object proposals has become a critical step in top-performing object detection systems [1, 2, 3], which helps reduce the search space of detection to a relatively small number of interesting regions [4]. Such reduction improves not only the computational efficiency but also the accuracy of detection methods thanks to much fewer background clutters. Early work of object proposal generation focuses on exploiting local image cues, including object contour [5], edge density [6] and over-segmentation [7, 8, 9]. It usually requires generating thousands of object proposals per image to achieve high recall rate and accurate localization in detection. More recently, learning-based methods have been proposed to refine an initial set of proposals or to directly generate them from images based on deep network features [10, 11, 12, 13]. In addition, 3D shape cues are learned from dense depth images for indoor scenes [14]. These new proposal generation methods generally further improve the quality of object proposals and lead to better object detection and localization performance.

Despite the progress, most of existing proposal generation approaches extract objectness cues from single modality and focus on low- or mid-level features. On the other hand, the spatial locations of object instances need to satisfy certain geometric/physical constraints and have close relations to their neighboring object classes, such as supporting relation and relative size. As such, incorporating geometric and semantic context cues can benefit the proposal generation and further improve their quality.

It has been widely acknowledged that global context plays an important role in object detection and recognition [15]. Several types of contextual information have been explored in the object detection literature, such as scene geometry [16], co-occurring object classes [17], and semantic scene labeling [18]. However, little attention has been paid to exploiting context information in the stage of object proposal generation. A notable exception is the recent work by Chen et al [3], which uses depth context to improve the object proposal generation. However, they focus on the class-dependent object proposals and use estimated ground plane to reduce their search space, which is restrictive for generic scene understanding.

In this work, we propose a novel object proposal generation pipeline, which exploits additional geometric and semantic context cues to improve the recall and localization accuracy of object proposals. To this end, we take a pair images from a stereo camera as input and start from a set of initial object proposals generated from applying the Edgebox method [6] to the left image. Our goal is to refine this set of proposals by re-ranking them and finetune their spatial locations based on a new set of object and context information.

Specifically, we consider the following three kinds of objectness cues. First, we use the noisy depth computed from the stero images to estimate a set of geometric features on each object candidate; second, we design a semantic context feature to describe the surrounding object class distribution, which is computed from a noisy semantic labeling; finally, we follow the Deepbox method [11] and extract a CNN feature from each object candidate. We then fuse these object and context cues to re-rank the initial object candidates. In particular, based on those features, we train a classifier to predict a new objectness score for each candidate, and regressors to adjust the location of its bounding box. Fig.1 illustrates the overview of our approach.

We evaluate our method on the KITTI dataset [19], one of the large-scale publicly available datasets with both stereo

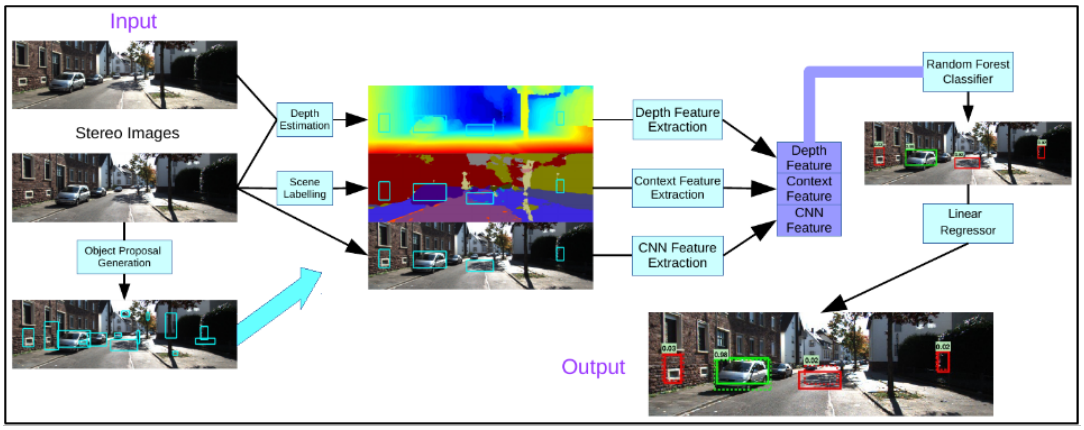

Fig. 1. Overview of our object proposal generation pipeline. The input are a pair of stereo images and an initial set of proposals. We extract three types of object and context cues, and use them to re-rank the proposals and refine their locations.

images and object annotation. We show that our method improves the quality of the initial object proposals significantly and achieves the state-of-the-art performance. Our main contributions are summarized as follows: 1) We propose a new pipeline for improving object proposals based on additional geometric and semantic context cues; 2) We design a set of geometric and semantic context features that can be efficiently computed (Section 2); 3) We systematically evaluate our method on the KITTI dataset and achieve the state-of-the-art recall rate with much fewer proposals (Section 3).

2. OUR APPROACH

We take as our system input a pair of stereo images and aim to generate a set of high-quality object proposals for its left image. Our approach consists of three stages, as illustrated in Fig. 1. We first generate a set of initial object proposals in the left image. Given the initial object proposals, we then compute three sets of object and context features for each object proposal, including its geometric properties, the CNN feature and a semantic context feature. Finally, we concatenate these features and train a classifier to re-rank as well as regressors to re-locate those initial candidates. We now introduce the details of each stage of our pipeline, focusing on the feature design and classifier plus regressor training.

2.1. Preprocessing

The preprocessing stage generates a set of initial object proposals, dense depth and semantic maps for computing context features in the next stage. For the initial object proposals, we choose the Edgebox algorithm [6] for its efficiency and good Intersection-Over-Union (IOU) quality. We use the disparity estimation method [20] to estimate the dense depth map and convert it into a point cloud representation according to the camera parameters. The semantic map is computed based on the SegNet system [21], although any deep Convnet based method can be used here. The SegNet is pre-trained on the CamVid dataset [22] and generates a pixel-level label map with 12 semantic classes, which are commonly seen in street scenes. We note that no object instance information is available from their outputs.

2.2. Object and Context Features

Given each initial object proposal, we compute three types of features to capture its appearance, shape and its geometric context, as well as the semantic context.

CNN Feature For each candidate bounding box, we adopt the CNN feature to encode the object appearance. Specifically, we extract the CNN feature in the same way as in the R-CNN method [1]. We normalize each bounding box into a size of 224×224 and apply the AlexNet [23] network. The network weights are pre-trained on the ImageNet [24] and fine-tuned on the VOC 2012 dataset [25]. We take the output from the layer fc6 as our CNN feature, which has 4096 dimensions.

Geometric Feature To incorporate geometric property of the object, we make use of the depth map estimated from the stereo images. We first segment out the subset of the point cloud using the bounding box associated with a proposal. The subset is used to compute a 12-dimensional feature to describe the object's geometric properties. Specifically, denoting the position of a 3D point as (x, y, z), we consider the following set of features, including *mean x, mean y, mean z, median x, median y, median z* of all points in the bounding box and the *x, y and z* of the center point, as well as the *width, height and depth span* of all points in the box.

Semantic Context We encode the semantic context of each object proposal by computing a semantic layout feature on the pixel-wise semantic label map. Specifically, each pixel is labeled into 12 classes: *sky, road, road marking, building, pavement, wall/fence, pole, vegetation, car, pedestrian, sign, cyclist*. We split the bounding-box into $n \times n$ cells (we use $n = 6$ in our experiment) on the label map. For each of those

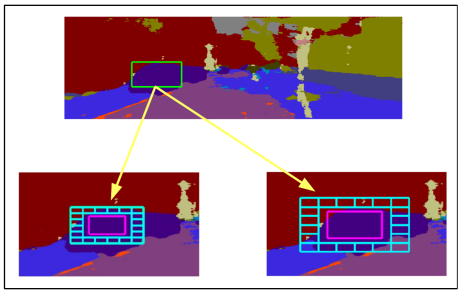

Fig. 2. The design of semantic context feature, which shows the partition of a bounding box for computing the label histogram. See text for details.

cells which are next to the boundaries($4n-4$ cells in total), we compute a label histogram. Besides that we also compute the label histogram of the inner box whose area is a quarter of the original bounding-box. In order to better capture context information, we enlarge the original bounding box by 1.5 times in terms of area and then compute the histograms in the same way as for the original bounding box. Finally, we concatenate these histograms computed from the original and the enlarged bounding box as the semantic context feature. Fig.2 shows an example of computing the semantic context feature.

2.3. Re-rank proposals

We concatenate all the features computed from Section 2.2 and re-rank all the initial object proposals based on these features. We adopt the random forest (RF) [26] as our classifier for its efficiency during test. To train the random forest classifier, we build our training dataset as follows. We treat the ground-truth bounding boxes and those proposals with ≥ 0.5 IoU overlap with a ground-truth box as positives. Those proposals with ≤ 0.4 IoU overlap with a ground-truth box are labeled as negatives. We use the held-out validation set to optimize the hyper-parameters in the RF classifier. Our RF classifier consists of 15 trees with a maximal depth of 20 and at least 2 leaf nodes. The RF generates a probability score for each proposal, which is used as the new objectness score.

2.4. Bounding box regression

Inspired by [1], we learn a bounding box regressor to fine-tune the location of each proposal. Note that our regressors are class-agnostic. We represent the bouning-box by its center coordinates, its width and height, $\{B_x, B_y, B_w, B_h\}$. The ground-truth box is denoted as $\{G_x, G_y, G_w, G_h\}$. We define the regression targets $\{T_x, T_y, T_w, T_h\}$ for the training pair (B, G) as follows,

$$T_x = (G_x - B_x)/B_w, \quad T_y = (G_y - B_y)/B_h \tag{1}$$

$$T_w = \log(G_w/B_w), \quad T_h = \log(G_h/B_h) \tag{2}$$

We learn four linear regressors with the same features as the RF classifier. For each regressor, we estimate the weights β by minimizing the regularized least squares objective:

$$\beta = \arg\min_{\widehat{\beta}} \sum_{i=1}^{N} (T^i - \widehat{\beta}^T f(B^i))^2 + \lambda \parallel \widehat{\beta} \parallel^2, \tag{3}$$

where $f(B^i)$ denotes the feature extracted for the bounding box B^i, and λ is the weight for the regularization term. For learning these regressors, we only use those proposals which have ≥ 0.5 IoU overlap with a ground-truth box.

3. EXPERIMENTS

We evaluate our approach on the KITTI object dataset [19], which consists of 7481 images with bounding box annotations. The object classes consist of $Cars$, $Pedestrains$ and $Cyclists$. Similar to the setup in [3], we split the dataset into three subsets: a training set of 3200 images, a validation set of 512 images and a test set of 3769 images. We report the results of object proposal generation and object detection task on the test set.

3.1. Object proposal generation

For object proposal generation, we employ the recall vs. number of proposals and the recall vs. IoU threshold as the evaluation metrics. For the recall vs. the number of proposals, we use 0.5 as the IoU threshold, above which a proposal is treated as recalled. For the recall vs. the IoU, we use top 100 and 1000 proposals to evaluate the performance.

We first compare our algorithm against the baseline method, Edgebox-50 [6], and the state-of-the-art, 3DOP [3]. Fig. 3 (left) shows the recall when varying the number of object proposals. We can see that our approach significantly improves the recall rate. With just 100 proposals, our approach improves the recall rate to a level above 90%, while 3DOP and EdgeBoxes only achieve 63% and 30% respectively. Furthermore, with recall rate 90%, our method uses only one tenth as many proposals as the 3DOP method, which leads to more efficient object detection. We can also see that the bounding box regression further improves the recall rate of our method. This implies that the depth cue also help refine the quality of the initial proposals.

We also show the recall rate when changing the IoU threshold with top 100 and 1,000 proposals in Fig. 3 (middle and right). We can see that our approach greatly outperforms the baseline and the state-of-the-art. Interestingly, the bounding box regression improves the proposals location precision significantly. We note that 3DOP uses the object size priors learned for each class, which are unavailable to our method.

3.2. Ablation Study

To understand the effectiveness of different features, we conduct the ablation study as follows. In the re-ranking stage, we use different groups of features to train the classifier. Fig. 4 (left) shows the recall rate curves with different combinations of our features. We can see that using the geometry features or

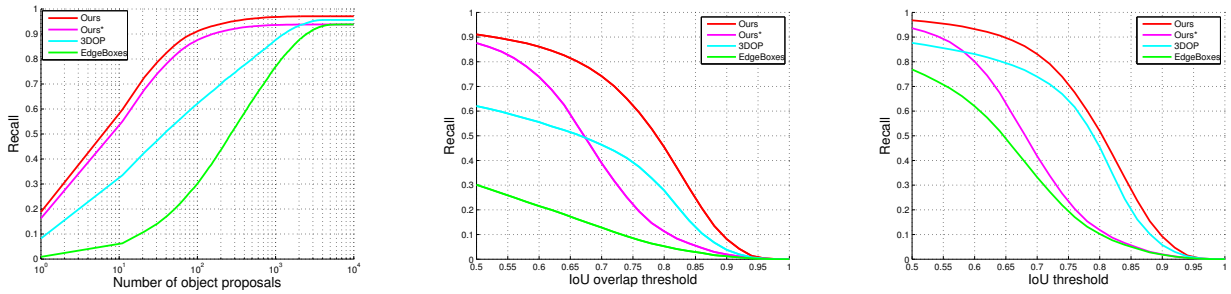

Fig. 3. Comparison of our approach to the baseline and the state-of-the-art (3DOP). 'Ours*' denotes our approach without the bounding box regression.**Left**: Recall vs. Number of proposals; **Middle**: Recall vs. IoU Threshold (100 proposals); **Right**: Recall vs. IoU Threshold (1000 proposals).

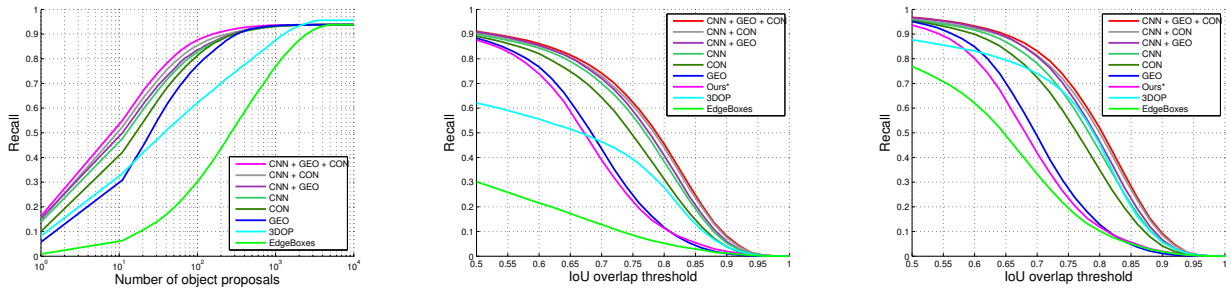

Fig. 4. Ablation study of our features on proposal re-ranking and bounding box regression. **Left**: Effectiveness of features on the object proposals re-ranking; **Middle**: Effectiveness of features on the bounding box regression (100); **Right**: Effectiveness of features on the bounding box regression (1000).

	Cars			Pedestrians			Cyclists		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
3DOP + Random Forest	45.74	37.79	32.48	51.62	45.57	**41.24**	**29.96**	**22.41**	**21.30**
Ours + Random Forest	**52.39**	**44.88**	**37.33**	**52.21**	**46.45**	41.02	23.51	21.84	20.59

Table 1. Average Precision (%) of object detection on the test subset with top 1,000 proposals. We use the class-agnostic version of 3DOP and our approach to generate the proposals respectively.

the semantic context feature alone can improve the recall rate greatly. All the features contribute to the final improvement of recall performance. We also apply the same study to the regression stage and show the results in Fig. 4 (middle for 100 proposals and right for 1000). We can see that the geometry features are not very effective in the bounding box regression, but the context feature is quite powerful. Both studies verify the strength of the CNN feature.

3.3. Object Detection

To demonstrate the benefit of our proposal generation method, we evaluate the performance of object detection task using our proposals. We train a set of object detectors based on a random forest classifier (20 trees with a maximal depth of 25 and at least 3 leaf nodes), which take the same feature set as in Sec 2.2. We compare the results using our proposals and the proposals from the class-agnostic version of 3DOP as the input to the detectors. Table 1 shows the average precision of two systems. Our proposals perform better than 3DOP's in the majority cases. For the category of cyclist, 3DOP uses the learned 3D size priors, which can help get more precise proposals, as it can be difficult to discriminate the pedestrians from the cyclists.

4. CONCLUSION

In this paper, we propose a new object proposal generation method for stereo images, which exploits additional geometric and semantic context cues. In addition to the CNN feature of proposals, we design geometric features based on depth map and a semantic context feature computed from pixel-level scene labeling. We train an efficient classifier to re-rank the initial object proposals, and learn a set of bounding box location regressors to fine-tune the position of the re-ranked object proposals. Experiments on the KITTI dataset show that our approach achieves high recall rate with a fraction of the initial proposals and outperforms the state-of-the-art.

5. REFERENCES

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Region-based convolutional networks for accurate object detection and segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2015.

[2] Ross B. Girshick, "Fast R-CNN," *CoRR*, vol. abs/1504.08083, 2015.

[3] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun, "3d object proposals for accurate object class detection," in *NIPS*, 2015.

[4] Jan Hosang, Rodrigo Benenson, Piotr Dollár, and Bernt Schiele, "What makes for effective detection proposals?," *arXiv preprint arXiv:1502.05082*, 2015.

[5] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *IEEE CVPR*, 2014.

[6] C Lawrence Zitnick and Piotr Dollár, "Edge boxes: Locating object proposals from edges," in *Computer Vision–ECCV 2014*, pp. 391–405. Springer, 2014.

[7] Joao Carreira and Cristian Sminchisescu, "Cpmc: Automatic object segmentation using constrained parametric min-cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 7, pp. 1312–1328, 2012.

[8] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[9] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik, "Multiscale combinatorial grouping," CVPR, 2014.

[10] Philipp Krähenbühl and Vladlen Koltun, "Learning to propose objects," in *CVPR*, 2015.

[11] Weicheng Kuo, Bharath Hariharan, and Jitendra Malik, "Deepbox: Learning objectness with convolutional networks," *arXiv preprint arXiv:1505.02146*, 2015.

[12] Amir Ghodrati, Ali Diba, Marco Pedersoli, Tinne Tuytelaars, and Luc Van Gool, "Deepproposal: Hunting objects by cascading deep convolutional layers," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2578–2586.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[14] Shuai Zheng, Victor Adrian Prisacariu, Melinos Averkiou, Ming-Ming Cheng, Niloy J Mitra, Jamie Shotton, Philip HS Torr, and Carsten Rother, "Object proposals estimation in depth image using compact 3d shape manifolds," in *Pattern Recognition*, pp. 196–208. Springer, 2015.

[15] Carolina Galleguillos and Serge Belongie, "Context based object categorization: A critical survey," *CVIU*, vol. 114, pp. 712–722, 2010.

[16] Derek Hoiem, Alexei a. Efros, and Martial Hebert, "Putting Objects in Perspective," *IJCV*, vol. 80, pp. 3–15, 2008.

[17] Chaitanya Desai, Deva Ramanan, and Charless Fowlkes, "Discriminative models for multi-class object layout," in *ICCV*, 2009.

[18] Joseph Tighe and Svetlana Lazebnik Marc Niethammer, "Scene parsing with object instances and occlusion ordering," in *CVPR*, 2014.

[19] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, p. 0278364913491297, 2013.

[20] Koichiro Yamaguchi, David McAllester, and Raquel Urtasun, "Efficient joint segmentation, occlusion labeling, stereo and flow estimation," in *Computer Vision–ECCV 2014*, pp. 756–771. Springer, 2014.

[21] Vijay Badrinarayanan, Ankur Handa, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," *arXiv preprint arXiv:1505.07293*, 2015.

[22] Gabriel J. Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla, "Segmentation and recognition using structure from motion point clouds," in *ECCV*, 2008.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[25] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, June 2010.

[26] Piotr Dollár, "Piotr's computer vision matlab toolbox (pmt)," http://vision.ucsd.edu/ pdollar/toolbox/doc/index.html.