PEDESTRIAN PROPOSAL GENERATION USING DEPTH-AWARE SCALE ESTIMATION

Kihong Park *Seungryong Kim* *Kwanghoon Sohn*

School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea
E-mail: khsohn@yonsei.ac.kr

ABSTRACT

In this work, we propose an efficient method that generates pedestrian proposals suitable for the autonomous vehicle. Our main intuition is that depth information provides an important cue to assign the scale of pedestrian proposals. Based on the observation that in a 3-D world coordinate the scales of pedestrians are almost similar, we formulate the scales of pedestrian patches by projecting 3-D models to an image plane with its corresponding depth. We also introduce a scale-aware binary description using both color and depth images. By using this descriptor, the regression models are trained to rank the pedestrian proposal candidates and adjust the proposal bounding boxes for an accurate localization. Our algorithm achieves significant performance gains compared to conventional proposal generation methods on the challenging KITTI dataset.

Index Terms— RGB-D object proposal, scale estimation, scale-invariant feature. binary feature selection, pedestrian detection

1. INTRODUCTION

Pedestrian detection has attracted a special attention as one of fundamental tasks for numerous real-world applications such as automatic driving and intelligent surveillance [1, 2, 3]. To detect a pedestrian in a scene, the pedestrian proposal generation step is essential to improve a detection performance by limiting pedestrian candidates in the scene while eliminating unreliable pedestrian candidates.

Although many methods have been proposed for that task [4, 5, 6], they have frequently encountered a scale ambiguity that hinders an optimal performance, which means that the scales of pedestrians in an image vary when pedestrians appear in a wide range of distances. Unlike object proposal estimation [7], this scale ambiguity is one of the most important issues in pedestrian proposal generation. To overcome this problem, most object proposal estimation methods attempt to exploit multi-scale process and scale prediction process [5, 8]. First of all, multi-scale process based approaches, adopted in many hand-crafted methods such as BING [5] and edge box (EB) [9], first resize an image to multiple quantized scales and then iteratively generate pedestrian proposals on each resized image. However, their computational complexity is proportion to the number of the scale levels, and false positives also highly increase during iteration process. Secondly, scale prediction process based approaches are to estimate a scale of each proposal explicitly and utilize this scale to detect pedestrian proposals. By leveraging the powerful learning capabilities of convolutional neural networks (CNNs), these methods have been popularly proposed and provided outstanding performances [8, 1]. However, when estimating the scale of a pedestrian within a low resolution, it cannot produce stable performances due to low discrimination power of CNNs on low-resolution feature maps [10]. The high complexity of CNN based methods is also a major hurdle to real-world applications when considering a limited hardware computing power.

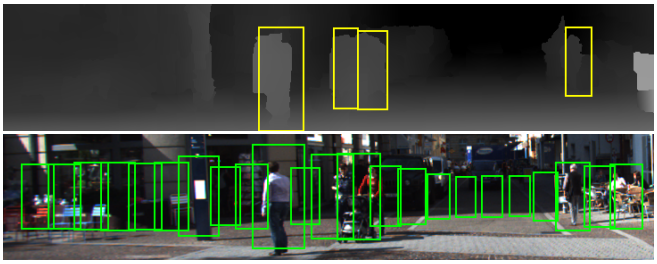

Fig. 1. Pedestrian scale estimation with a depth value. Even if two pedestrians have different scales in an image, we assign the scale of pedestrians using depth information and detect pedestrian proposal candidates reliably.

Recently, some researchers have tried to estimate more reliable proposals by utilizing depth information to supplement the data provided in color images [11, 6, 10]. While a color image is sensitive to complex intra-class variations, such as different colors or illumination variations, in pedestrian candidates, its corresponding depth image provides simple but intuitive geometric information. Thus, the performance of pedestrian proposal generation can be boosted when color and depth images are jointly used in a synergistic manner. Multiscale combinatorial grouping (MCG) [11] generated proposal candidates by combining features from the color and depth images. 3D object proposal (3DOP) [6] also estimated proposals in a 3-D point cloud domain created using color and depth information. Zhang *et. al.* [10] computed and fused three types of features to capture appearance, geometric, and semantic context information simultaneously. Thanks to the high accurate scale estimation performance, the pedestrian proposal detection accuracy combined with these methods can be more enhanced than when color information is considered alone in existing methods [12]. Since, however, their feature descriptions need huge complexity processes such as multi-scale segmentation [11] or CNN activations [6], they cannot be directly applied to a real vehicle system.

In this paper, we introduce an efficient pedestrian proposal estimation algorithm that incorporates a scale estimation process using depth information in a context of the pedestrian detection. Our key-ingredient is to leverage depth information to determine the optimal scale of pedestrians. We observe that in a 3-D world coordinate, the scales of the pedestrians are almost similar although they vary in an image plane according to the distance from the camera, as shown in Fig. 1. Based on this observation, we propose a pedestrian proposal candidate generation that assigns the pedestrian scale in an image domain from the representative pedestrian scale in a 3-D world coordinate. We also propose a scale-invariant binary description scheme and a feature selection technique to construct robust descriptor efficiently. With this binary descriptor, the regression models are trained to rank and re-localize bounding boxes. With the accurate scale estimation and scale-invariant binary feature, the proposed algorithm solves the scale ambiguity and shows state-of-the-art proposal gen-

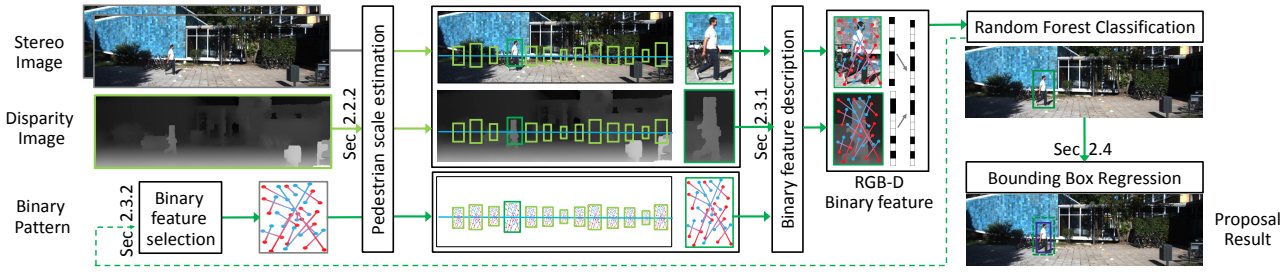

Fig. 2. Overall framework of proposed method.

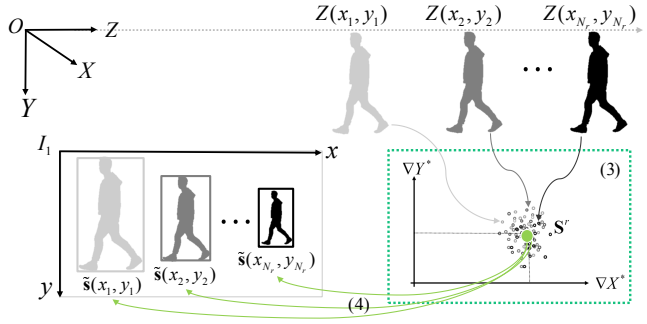

Fig. 3. Projection of the representative pedestrian scale in a 3-D world coordinates into an image plane.

eration performance compared to conventional proposal generation methods on KITTI dataset [13].

2. PROPOSED METHOD

2.1. Notation and Overview

We represent each pedestrian proposal with a rectangular bounding box which is parametrized by a tuple $[x, y, \nabla x, \nabla y]^T$, where (x, y) is the center point of the 2-D box. $\mathbf{s} = [\nabla x, \nabla y]^T$ denotes the horizontal and vertical diameters, which will be termed "pedestrian scale". Different from traditional approaches that leverage an image I only, the proposed method explicitly incorporates scale information, obtained from a disparity map D, into the proposal generation. Our key observation is that in a 3-D world coordinate, the scales of pedestrian are almost similar although they vary in an image plane according to the distance from the camera, i.e., depth (see Fig. 3). Thus, when the representative pedestrian scale $\mathbf{S}_r$ in a 3-D world coordinate is determined, the corresponding scale $\mathbf{s}$ in an image plane also can be estimated using depth information.

We first estimate the disparity map D between stereo pair I and I' using the method [14], and transform it into the depth Z. With the pedestrian scale consistency assumption in a 3-D world coordinate, we define the representative pedestrian scale $\mathbf{S}_r$ in a 3-D world coordinate by using the ground-truth pedestrian box and its corresponding depth. After that, we generate the pedestrian proposal candidates by assigning the pedestrian scale $\mathbf{s}$ for each pixel from the representative pedestrian scale $\mathbf{S}_r$ according to its corresponding depth Z. Finally a random forest (RF) classifier [15] is trained with binary feature extractor [16], followed by the bounding box regression [17]. The overall framework of proposed method is illustrated in Fig. 2.

2.2. Proposal Candidate Generation Using Depth Information

2.2.1. *Preliminaries*

Formally the disparity map can be transformed into depth such that $Z(x, y) = fB/D(x, y)$, where f is the focal length and B is the baseline of a stereo camera. Using depth information, a point (x, y) in the image plane can be transformed into the 3-D world coordinate $\mathbf{P}(x, y) = [X, Y, Z]$ or vice versa:

$$\mathbf{P}(x, y) = Z(x, y)K^{-1}[x, y, 1]^T, \tag{1}$$

$$[x, y, 1]^T = 1/Z(x, y)K\mathbf{P}(x, y), \tag{2}$$

where $K = \text{diag}([f, f, 1])$ is the intrinsic camera matrix. Note that we do not consider skew parameter for the simplicity of notation.

2.2.2. *Pedestrian scale estimation*

Our approach is primarily based on the scale-invariance property of pedestrians in the 3-D world coordinate. Moreover, the pedestrian is located on a front-parallel space, which means that depth of a pedestrian is constant. To verify this, we estimate the distributions of the ground-truth pedestrian scale $\mathbf{S}^* = [\nabla X^*, \nabla Y^*]^T$ in the 3-D world coordinate using 2446 user-annotations in the KITTI dataset [13]. As shown in Fig. 3, the distributions of $\mathbf{S}^*$ are concentrated on a single point, which implies that most pedestrians can be modeled with a single representative scale in the 3-D world coordinate. Note that the standard deviations of each distribution $(\sigma_{\nabla X^*}, \sigma_{\nabla Y^*})$ are 0.0029 and 0.0034, respectively. To estimate the representative scale $\mathbf{S}_r$ for the pedestrian in the 3-D world coordinate, we adopt a simple strategy that averages the ground-truth scale using KITTI training dataset [13]. Specifically, for all ground truth bounding box, $\mathbf{S}_r$ can be estimated as follows:

$$\mathbf{S}_r = \frac{1}{N_r}[\sum_{l=1}^{N_r} \nabla X_l^*, \sum_{l=1}^{N_r} \nabla Y_l^*]^T \tag{3}$$

where l and N_r denote the index and the total number of training dataset.

For the testing image and each point (x, y), the pedestrian scale $\tilde{\mathbf{s}}(x, y)$ can be now predicted by projecting the representative scale $\mathbf{S}_r$ into the image coordinate:

$$[\tilde{\mathbf{s}}(x, y), 0]^T = \frac{1}{Z(x, y)} K[\mathbf{S}_r^T, 0]^T. \tag{4}$$

That is, the estimated pedestrian scale in the image plane is inversely proportional to the corresponding depth information. Bounding box candidates of each point (x, y) are then defined as follows:

$$\mathbf{c} = [x, y, \tilde{\mathbf{s}}(x, y)]^T. \tag{5}$$

2.3. Binary Feature with Scale-aware Pattern Generation

For bounding box candidates, the pedestrian proposals can be determined through the feature description on the regions and classifier. To effectively describe the bounding box candidates, we adopt a binary descriptor due to its computational efficiency and low memory consumption. Since in the binary descriptor, binary sampling patterns are fixed for all pixels in an image, we transform the sampling

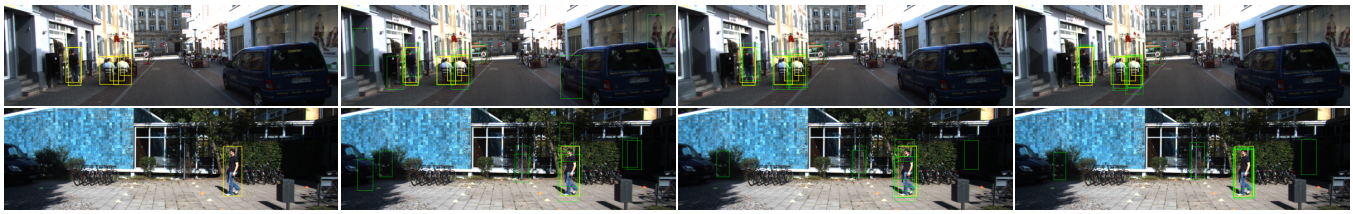

Fig. 4. Component contribution on our pedestrian proposal generation: (from left to right) Ground truth, pedestrian proposal detections using our method without a feature selection and bounding box regression, without bounding box regression, and using our final method.

patterns according to its corresponding scale. Since such a process does not need an additional computational time, the binary descriptor can be still built in the image domain very efficiently. Since, however, this simple descriptor cannot handle intra-class variations due to the variety of pedestrian patches in color, lighting, backgrounds, and occlusion, we generate a robust descriptor by fusing color and depth features simultaneously.

2.3.1. Binary feature description

The binary descriptor is defined with pairwise intensity comparisons sampled on binary sampling patterns. It simply creates a bit vector with a comparison pattern as

$$R(I(x,y);u_k,v_k) := \begin{cases} 1 & \text{if } I(x,y;u_k) < I(x,y;v_k) \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

where u_k and v_k are k-th sampling pattern for $k \in \{1,...,N_p\}$ with the number of sampling patterns N_p. As in [16], it shows high discriminative power with low cost.

To describe each bounding box candidate with the scale $\tilde{\mathbf{s}}(x,y)$, we resize the sampling pattern (u_k, v_k) such that $(\tilde{u}_k, \tilde{v}_k) = (u_k \cdot \tilde{\mathbf{s}}(x,y), v_k \cdot \tilde{\mathbf{s}}(x,y))$. Thus, binary descriptor for each bounding box candidate can be described such that

$$B_I(x,y) := \sum_{k=1}^{N_p} 2^{k-1} R\left(I(x,y);\tilde{u}_k,\tilde{v}_k\right), \quad (7)$$

$$B_D(x,y) := \sum_{k=1}^{N_p} 2^{k-1} R\left(D(x,y);\tilde{u}_k,\tilde{v}_k\right), \quad (8)$$

$$B(x,y) = B_I(x,y) + 2^{N_p} B_D(x,y). \quad (9)$$

where $B^I(x,y)$, $B^D(x,y)$ are the binary descriptors extracted from the color and depth image. A final binary descriptor $B(x,y)$ is constructed by concatenating $B_I(x,y)$ and $B_D(x,y)$.

2.3.2. Binary feature selection

The performance of pedestrian proposal detection depends on the quality of binary features. Moreover, the robustness of binary features depends on the sampling patterns. Even though there are several binary sampling pattern selection methods [18, 19] that provide a limited performance, we propose a binary sampling pattern selection scheme based on the RF [15] regression. Specifically, to construct the regression model with the optimum feature set, the RF [15] training is performed twice. We first train with the binary descriptor of N_f dimensions and select nodes of first N_l layers from N_t trees. Their predictors are used as a new N_s-dimension descriptor because these predictors are determined to maximize the information function. A final regression model is then obtained by retraining the RF [15] with this binary feature set.

2.4. Bounding Box Regression

Since our method is based on the pre-estimated optimal pedestrian scale, there might be approximation errors. To reduce such effects, inspired by bounding-box regression of [17], we train a regression model with the estimated binary descriptor to obtain more accurate proposal. Specifically, the input to our bounding box regression model is a set of N_r ground truth training pairs $\{(\mathbf{c}_l, \mathbf{g}_l)\}_{l=1,...N_r}$, where $\mathbf{c}_l = [x_l, y_l, \nabla x_l, \nabla y_l]^T$ specifies the pixel coordinates of the center, width and height of the initial proposal. We will drop the subscript l for convenience. The ground-truth box $\mathbf{g}$ is denoted in same way $\mathbf{g} = [x^*, y^*, \nabla x^*, \nabla y^*]^T$. The transformation $\mathbf{t} = [t_x, t_y, t_{\nabla x}, t_{\nabla y}]^T$ between $\mathbf{c}$ and $\mathbf{g}$ is trained as

$$\begin{aligned} t_x &= (x^* - x)/\nabla x, \;\; t_y = (y^* - y)/\nabla y, \\ t_{\nabla x} &= log(\nabla x^*/\nabla x), \;\; t_{\nabla y} = log(\nabla y^*/\nabla y). \end{aligned} \quad (10)$$

We learn four linear regression model weights with the binary descriptor by optimizing the regularized least squares objective.

3. EXPERIMENTAL RESULT

Our pedestrian proposal detection method was implemented with RF [15] in C++ on Intel Core i7-4770 CPU at 3.40 GHz, and measured the runtime on a single CPU core. We set $\{N_t, N_l, N_f, N_s\} = \{16, 5, 10000, 496\}$ for the feature selection.

We evaluated our method compared to state-of-the-art methods, such as SS [12], EB [9], region proposal networks (RPN) [8], MCG [11], 3DOP [6], Fusion-DPM [20], region-based convolutional networks (R-CNN) [21], and multiveiw random forest (MV-RGBD-RF) [22], on the KITTI object dataset [13], which has 7,481 training and 7,518 testing images. Since the ground-truth pedestrian proposals are not available in the testing set, the training set was manually partitioned into 3,740 training images and 3,741 validation images. Note that it is guaranteed that training and validation set do not come from the same video sequence.

3.1. Proposal Detection Evaluation

3.1.1. Component contribution analysis

In this section, we analyzed the performance gain of key-components in our method, including the feature selection and the bounding box regression. Fig. 4 shows the top 10 pedestrian proposals detected by our method using each component. As expected, the feature selection enhanced an accuracy of ranking because this component selects optimum feature set among the large number of feature candidates. The bounding box regression improved IoU overlaps. Even if we estimated the accurate pedestrian scales, various factors, such as body type, posture, and depth error, induce the scale variations, reducing the pedestrian detection performance. The bounding box regression corrected the errors from small scale differences.

3.1.2. Comparison with the state-of-the-art algorithms

Following qualitative experiments reported in [23, 6, 10], we measured the performance of proposal algorithms such as SS [12], EB [9], MCG [11], and 3DOP [6] including ours as shown in Fig. 5.

Figure 5 (a) shows recall rates as a function of the number of proposals. While the 3DOP [6], the state-of-the-art RGB-D proposal

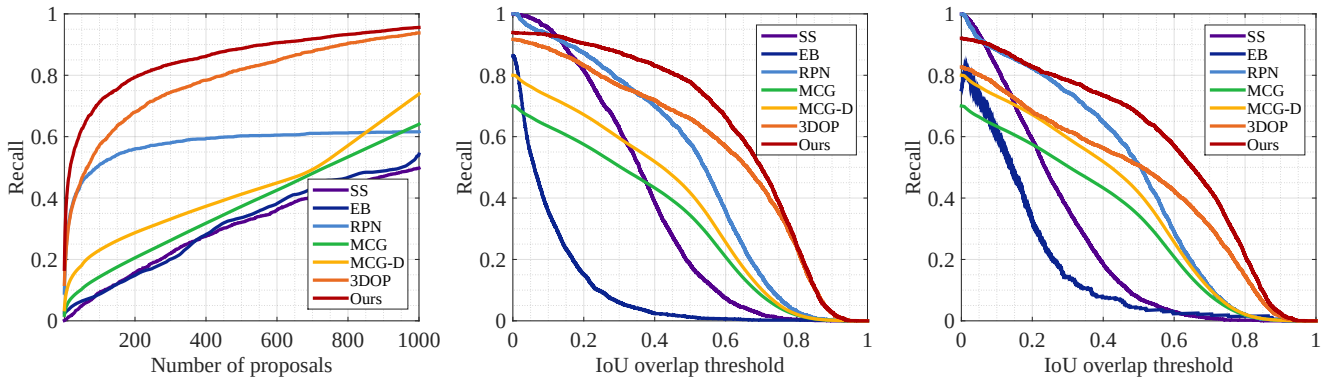

Fig. 5. Quantitative evluation of our pedestrian proposal estimation with the state-of-the-art algorithms: (a) Recall vs. number of proposals, (b) Recall vs. IoU Threshold (300 proposals), (c) Recall vs. IoU Threshold (100 proposals).

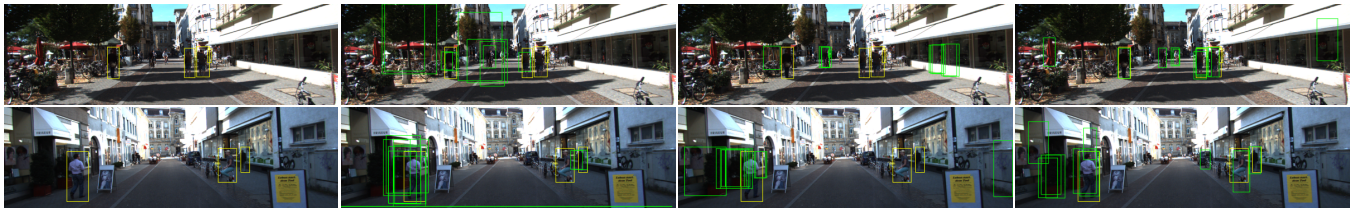

Fig. 6. Quantitative evaluation of our pedestrian proposal estimation with the state-of-the-art algorithms with visualization of top 10 pedestrian proposals: (from left to right) Ground truth, RPN [8], 3DOP [6], and proposed algorithm.

Table 1. Average precision (%) of object detection on the test subset with top 1,000 proposals.

Methods	Easy	Moderate	Hard	Avg.
Fusion-DPM [20]	59.51	46.67	42.05	49.41
R-CNN [21]	61.61	50.13	44.79	52.17
MV-RGBD-RF [22]	70.21	54.56	51.25	58.67
3DOP [6]	81.78	67.47	64.70	71.32
Ours	**84.37**	**69.18**	**67.50**	**73.68**

generation algorithm, requires 500 proposals to achieve 70 percent recall, our algorithm only requires 150 proposals to achieve the same recall. Moreover, Figure 5 (b), (c) show recall rates for 300 and 100 proposals as a function of IoU threshold. We achieved 5% and 9% improvements in average recall rates than the 3DOP [6]. The proposed method ensures reliable results even with a small number of proposals. This property was also presented in Fig. 6 that visualize only top 10 proposals, and the proposed method detects more reliable pedestrians similar to ground-truth than the comparison algorithms. RPN [8] fails to estimate accurate scale and thus shows low recall rate even if it achieves the high binary classification accuracy between pedestrians and backgrounds. 3DOP [6] shows the low regression performance at the ranking process. On the other hand, our algorithm accurately estimates the scales and the ranks of bounding boxes based on the scale-invariant binary description. It proves that the concept of the depth based scale estimation and the scale-invariant description contributes significantly to improving the accuracy of the proposal generation. Table 2. shows the computational complexity of the state-of-the-art algorithms. Our algorithm is efficient than other algorithm.

Table 2. Computation time of the-state-of-the-art proposal methods for handling an image size 1242×375.

Methods	SS [12]	EB [9]	MCG [11]	3DOP [6]	Ours
Time (sec.)	15.39	1.52	156.73	1.20	**0.63**

3.2. Pedestrian Detection Evaluation

To demonstrate the performance gain of our pedestrian proposal algorithm for pedestrian detection framework, we scored pedestrian proposal detection results using the detector of FAST R-CNN [17]. The results were denoted in Table 1. Our approach outperformed the state-of-the-art algorithms. We achieved 1.71% improvement in Average Precision (AP) for pedestrians.

4. CONCLUSION

We proposed the efficient proposal generation algorithm for pedestrian detection. With the scale-invariance property of pedestrians in the 3-D world coordinate, we utilized depth information to assign the scale of pedestrian bounding box candidates. To describe the pedestrian candidates effectively and efficiently, we adopted the binary feature descriptor to rank the pedestrian candidates and adjust the pedestrian bounding boxes. Our approach has shown outstanding performances compared to existing state-of-the-art pedestrian proposal algorithms on the challenging KITTI benchmark [13]. Combined with detection network of FAST R-CNN [17], our method definitely outperformed conventional pedestrian detection algorithms.

5. ACKNOWLEDGMENTS

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP)(No.2016-0-00197)

6. REFERENCES

[1] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," *In Proc. of ECCV*, 2016.

[2] X. Wang, M. Wang, and W. Li, "Scene-specific pedestrian detection for static video surveillance," *IEEE Trans. PAMI*, vol. 36, no. 2, pp. 361–374, 2013.

[3] H. Choi, S. Kim, K. Park, and K. Sohn, "Multi-spectral pedestrian detection based on accumulated object proposal with fully convolution network," *In Proc. of ICPR*, 2016.

[4] K. Kim, C. Oh, and K. Sohn, "Non-parametric human segmentation using support vector machine," *In Proc. of BMVC*, 2015.

[5] M. Cheng, N. J. Mitra, Z. Zhang, W. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," *In Proc. of CVPR*, 2014.

[6] X. Chen, K. Kunda, Y. Zhu, and A. Berneshawi, "3d object proposals for accurate object class detection," *In Proc. of NIPS*, 2014.

[7] R. S. Pahwa, J. Lu, N. Jiang, T. T. Ng, and M. N. Do, "Locating 3d object proposals: A depth-based online approach," *IEEE Trans. CSVT*, 2017.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *In Proc. of NIPS*, 2015.

[9] C. L. Zitnick and P. Dollar, "Edge boxes: Locating object proposals from edges," *In Proc. of ECCV*, 2014.

[10] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster r-cnn doing well for pedestrian detection?," *In Proc. of ECCV*, 2016.

[11] P. Arbelaez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," *In Proc. of CVPR*, 2014.

[12] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013.

[13] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," *In Proc. of CVPR*, 2012.

[14] K. Yamaguchi and R. McAllester, D. Urtasun, "Efficient joint segmentation, occlusion labeling, stereo and flow estimation," *In Proc. of ECCV*, 2014.

[15] L. Breiman, "Random forest," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," *In Proc. of ECCV*, 2010.

[17] R. Girshic, J. Donahue, T. Darrel, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[18] F. Fleuret, "Fast binary feature selection with conditional mutual information," *JMLR*, vol. 5, no. 1, pp. 1531–1555, 2004.

[19] A. Asaithambi, V. Valev, A. Krzyzak, and V. Zeljkovic, "A new approach for binary feature selection and combining classifiers," *In Proc. of HPCS*, 2014.

[20] C. Premebida, J. Carreira, J. Batista, and U. Nunes, "Pedestrian detection combining rgb and dense lidar data," *In Proc. of IROS*, 2014.

[21] Jan. Hosang, M. Omran, R. Benenson, and S. Bernt, "Taking a deeper look at pedestrians," *In Proc. of CVPR*, 2015.

[22] A. Gonzalez, G. Villalonga, J. Xu, D. Vazquez, J. Amores, and A. Lopez, "Multiview random forest of local experts combining rgb and lidar data for pedestrian detection," *In Proc. of IV*, 2015.

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