2019 IEEE Intelligent Vehicles Symposium (IV)
Paris, France. June 9-12, 2019

Enhanced Object Detection in Bird's Eye View Using 3D Global Context Inferred From Lidar Point Data

Yecheol Kim, Jaekyum Kim, Junho Koh, and Jun Won Choi

Abstract— **In this paper, we present a new deep neural network architecture, which detects objects in bird's eye view (BEV) using Lidar sensor data in autonomous driving scenarios. The key idea of the proposed method is to improve the accuracy of the object detection by exploiting the 3D global context provided by the whole set of Lidar points. The overall structure of the proposed method consists of two parts: 1) the detection core network (DetNet) and 2) the context extraction network (ConNet). First, the DetNet generates the BEV representation by projecting the Lidar points into the BEV plane and applies the CNN to extract the feature maps locally activated on the objects. The ConNet directly processes the whole set of the Lidar points to produce the $1 \times 1 \times k$ feature vector capturing the 3D geometrical structure of the surrounding in the global scale. The context vector produced by the ConNet is concatenated to each pixel of the feature maps obtained by the DetNet. The combined feature maps are used to regress the oriented bounding box and identify the category of the object. The experiments evaluated on the public KITTI dataset show that the use of the context feature offers the significant performance gain over the baseline and the proposed object detector achieves the competitive performance as compared to the state of the art 3D object detectors.**

I. INTRODUCTION

Three dimensional (3D) object detection refers to a task of obtaining the information of the 3D box coordinate and the category of objects from the sensor data. In autonomous driving, 3D object detection provides the useful information on the dynamically changing environments around the ego-vehicle. 3D object detection can also be conducted in Bird's eye view (BEV) domain, where the bounding box containing the object is represented in an elevated view of the object from above. Though object detection in BEV does not provide full 3D information on the objects, it offers the sufficient information to conduct path planning and control for autonomous driving. From now on, we will refer to the task of detecting the objects in BEV domain as BEV object detection. Various types of sensors can be used for 3D object detection. Though the camera sensor provides the rich information such as shape, color, and texture useful to detect the object, it has the limited capability in capturing 3D geometric information needed for 3D object detection. In particular, the Lidar sensor offers accurate ranging information by transmitting a laser pulse and measuring the reflected pulses. By scanning 3D region of interest, the Lidar sensor can generate the collection of the 3D points of coordinate, called the *point cloud*. Such point cloud data provides the 3D geometrical structure on the surrounding, which can be used for object detection.

Recently, we have seen remarkable advance in computer vision technique owing to the emergence of the deep neural network (DNN). One of the DNN architecture, called convolutional neural network (CNN), extracts the features from the data of a 2D grid topology (e.g., RGB image) by repeatedly applying the convolutional operation followed by the nonlinear function. By training the CNN model with a number of training data examples, high-level abstract features can be successfully extracted from the data. Such high level features can be used to achieve great performance for a variety of challenging computer vision tasks.

The CNN has also been applied to achieve significant performance improvement in the object detection task. So far, various CNN-based object detectors have been proposed to detect the objects from the 2D camera images [1]–[3]. Recently, the DNN architecture has also been developed to detect the objects from the 3D Lidar points. Since the Lidar sensor offers accurate ranging information, the object detectors using the Lidar points are shown to achieve good performance in 3D object detection. The point data acquired by the Lidar sensor has different structure from the camera image. The Lidar data contains the un-ordered set of points. The distribution of the Lidar data is irregular, discontinuous and sparse in 3D space, which makes it hard to apply the CNN model without proper modification. There are two ways to generate the features from the Lidar points. The first approach is to project the 3D Lidar points onto the appropriately chosen 2D planes (e.g. front view or BEV domains) and apply the CNN model to find the features used for 3D object detection. The second approach is to employ the neural network specifically designed to process the Lidar points directly. The popular neural network designed for this purpose is the PointNet [4], [5], which applies multi-layer perceptron (MLP) and the max-pooling operation to produce the features.

We expect that the contextual cues that can be inferred from the background of the object can be used to improve the performance of the object detection. For example, the geometrical structure captured in the road, guard rails, curb, global traffic scenes would help detecting the objects in the driving scenarios though the baseline object detectors tend to rely only on the locally activated features. In this paper, we investigate how to design an effective 3D object detector which can effectively exploit the contextual cue captured in the whole scan of Lidar points. In fact, it has been revealed

Yecheol Kim, Jaekyum Kim, Junho Koh, and Jun Won Choi are with Dept. of Electrical Engineering, Hanyang University, Seoul 04763, Korea.
Corresponding Author: Jun Won Choi
Email: {yckim, jkkim, jhkoh}@spa.hanyang.ac.kr, junwchoi@hanyang.ac.kr

978-1-7281-0560-4/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: Hanyang University. Downloaded on August 24,2023 at 05:00:46 UTC from IEEE Xplore. Restrictions apply.

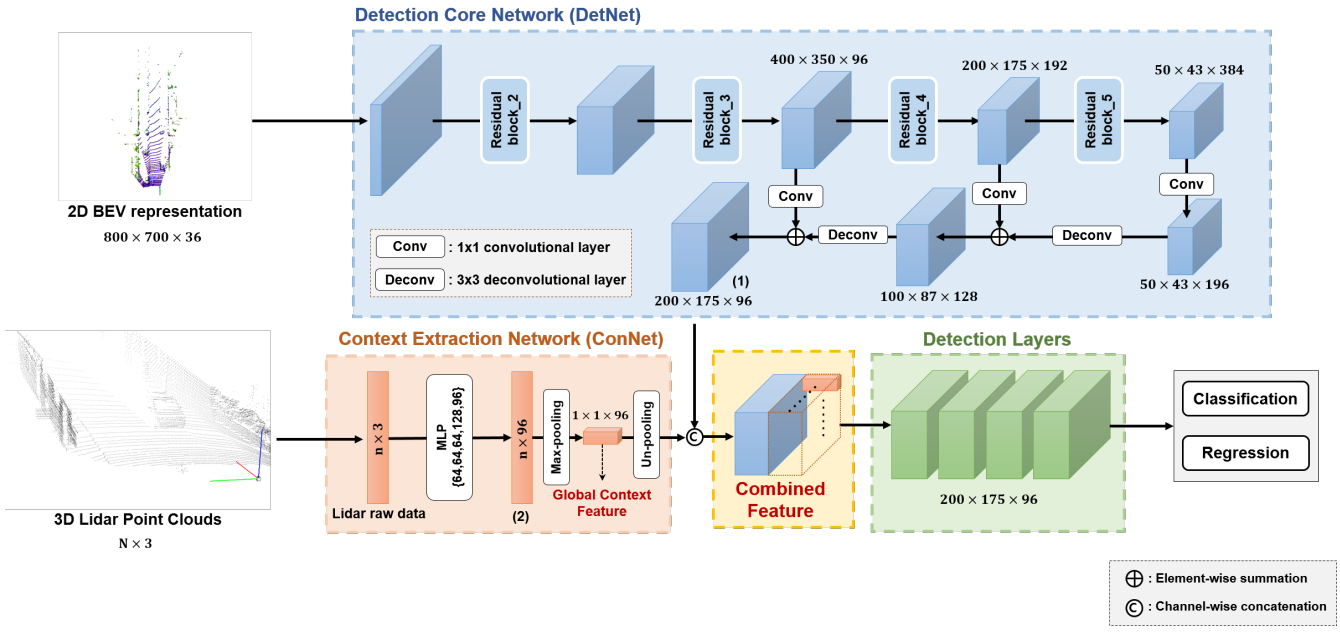


Fig. 1. Structure of the proposed BEV object detector

in [6] that such contextual information can be obtained from the HD map and the performance of the object detector can be improved using the context features.

In this paper, we propose a new BEV object detector which uses the 3D global context inferred from the scan of the Lidar points. Fig. 1 depicts the structure of the proposed object detector. Our method consists of two main parts; 1) the detection core network (DetNet) and 2) the context extraction network (ConNet). The DetNet takes the BEV representation of the Lidar points obtained by projecting them into the BEV plane and produces the feature maps for detecting the objects in BEV domain. On the other hand, the ConNet is designed to extract the contextual cues on the surrounding from the entire Lidar point set. The proposed ConNet employs the PointNet proposed in [4], which directly processes the raw Lidar points to produce the $1 \times 1 \times k$ feature vector called *context vector*. Such context vector exhibits the global scale information on the environment and is concatenated to each pixel of the feature maps obtained from the DetNet. The concatenated feature map is used to classify the object and regress the oriented bounding box. We train both DetNet and ConNet at the same time in an end to end fashion such that the ConNet learns to produce the complementary feature for improving the quality of the final feature maps. Our experiments conducted on the public KITTI benchmark [7] show that the global context generated by the ConNet offers significant performance improvement over the baseline BEV object detector. We also show that the proposed object detector achieves better performance than the state of the art BEV object detectors (top ranked on KITTI BEV object detection benchmark).

II. Related Works

In this section, we briefly review the previous 3D object detection methods, which use the Lidar points as an input. We specifically review two types of the DNN architectures for extracting the features from the 3D Lidar points.

A. 2D representation-based Object Detection

The MV3D [8] is the early work which tried to construct the 2D multi-view representation of the Lidar points (i.e., in front view and BEV) and apply the CNN to perform the 3D object detection. While the MV3D achieved impressive performance in 3D object detection, the computation time was burdensome for real-time applications. PIXOR [9] has reduced the complexity of the MV3D by developing the scheme that detects the objects only from the 2D BEV representation using the proposal-free and single-stage CNN architecture. The HDNET [6] improves the object detection performance significantly employing the map prediction module for extracting the geometric and semantic information from the HD map. The object detectors proposed in [10], [11] use both the Camera image and Lidar data for 3D object detection. The AVOD [10] fuses the Lidar BEV data and camera front data at the intermediate convolution layer to propose the 3D anchor boxes. The ContFuse [11] proposes the effective fusion architecture that transforms the front camera view features into those in BEV through some interpolation network.

B. PointNet-based Object Detection

Most object detectors under this category are built upon the PointNet [4] which was first proposed to extract the features from the Lidar points directly. The PointNet [4] and the PointNet++ [5] keep applying the MLP followed by the max-pooling to generate higher level features from the raw

Authorized licensed use limited to: Hanyang University. Downloaded on August 24,2023 at 05:00:46 UTC from IEEE Xplore. Restrictions apply.

Lidar points. The PointRCNN [12] employs the PointNet to segment the Lidar points in the foreground in the first step and then refines the 3D region proposals in the second step. The approaches proposed in [13]–[15] generate the 3D region proposals applying CNN to the camera image and then apply the PointNet to identify the object based on the Lidar points in the region of interest.

III. Proposed BEV Object Detection

In this paper, we present the details on the proposed BEV object detector.

A. Representation of Input Data

The Lidar sensor collects the 3D coordinate (x,y,z) and reflectivity r for each laser pulse, generating the point clouds around the ego-vehicle. The input to the DetNet is the 2D image constructed by projecting the Lidar points on the BEV plane. Specifically, we discretize the 3D point sets in the 3D region of $l_x \times l_y \times l_z$ into the 3D voxel grid of size $r_x \times r_y \times r_z$. In our setup, we discretize the region of $[0,70] \times [-40,40] \times [-2.5,1]$ meters at the resolution of 0.1 meters. As a result, the shape of the 2D BEV image becomes $700 \times 800 \times 36$. The pixel value of each voxel element is filled with the height and reflectivity values of the point with the maximum height among all points contained in each voxel element. Such discretized data is transformed into the $r_x \times r_y$ 2D image with the (r_z+1) channels (including r_z height channels and one reflectivity channel). We normalize all pixel values to be within $[0,1]$. On the other hand, the whole n 3D raw Lidar points are directly fed into the ConNet in the shape of $n \times 3$ matrix (i,e., (x,y,z) coordinate for each row).

B. Structure of the Proposed Method

1) Overall Structure: The whole structure of the proposed object detector is illustrated in Fig. 1. The DetNet takes the 700×800 BEV representation as an input and apply the CNN to generate the feature maps that retain the spatial information. Following the structure of the PIXOR network [9], we use the backbone network consisting of the four residual blocks [16] having $[3,6,6,4]$ layers for each. Then, the feature pyramid network (FPN) [17] is followed to increase the size of the feature maps while preserving the semantic depth of the features. Note that the size of the final feature maps at the output of the FPN is four times less than that of the original input. While the work in [9] directly regresses the box location without using the anchor box, our DetNet employs two anchor boxes for each pixel to achieve better accuracy at the expense of higher complexity.

In order to extract the 3D global context from the Lidar points, the ConNet takes the vector of the whole raw Lidar points (obtained from a single scan) as an input. Following the structure of the PointNet [4], the ConNet applies the MLP of the five layers $[64,64,64,128,96]$ and the max-pooling to the point vector and produces the $1 \times 1 \times 96$ feature vector. We concatenate such context vector to each pixel of the feature maps from the DetNet. (see Fig. 1.) The concatenated feature maps are finally fed into four additional 3×3 convolutional layers to produce the final score for the box regression and the object classification. The size of the channels is set to 96 for all extra convolutional layers. Our structure guides the global context feature produced by the ConNet to interact with the local feature obtained by the DetNet to yield better feature maps for BEV object detection.

2) Anchor boxes: The anchor box provides the reference for the oriented bounding box such that the residual set with respect to the reference are estimated by the regression model [1], [3]. Since the scale of the objects tends to be similar in BEV, we consider only two anchor boxes with two different orientations $\{0,\pi/2\}$. The anchor boxes have the area of 650 with the aspect ratio $2.3:1$ determined by averaging the size of ground truth bounding box for all vehicle objects in KITTI dataset. Each anchor is associated with K dimensional one-hot vector containing the classification confidence for K object classes and six regression parameters containing the center location (cx,cy), the size (w,h) and the orientation $(\cos 2\theta, \sin 2\theta)$. Note that the rotation angle θ of the bounding box is in the range $[-\pi/2,\pi/2]$ which disables distinguishing the front and back ends of the objects.

3) Loss Function: We adopt the multi-task loss used in [9] to train the proposed BEV detection network. In order to cope with class imbalance issue, we use the focal loss with the same hyper-parameter used in [3]. The focal loss for the classification task and smooth $\ell 1$ loss for the regression task are used. The total loss is the sum of the focal loss computed over all anchors and the smooth $\ell 1$ loss computed over all positive anchors. We consider the anchor to be positive if the intersection-over-union (IoU) is in $[0.5,1]$, and negative if it is in $[0,0.5)$.

IV. Experiments

In this section, we evaluate the performance of the proposed BEV object detection method using the public KITTI dataset [7].

A. Dataset

In the KITT dataset, the 3D Lidar point cloud data is collected using a Velodyne HDL-64E laser scanner. The KITTI dataset consists of 7,481 images for training and 7,518 images for test. Since the test labels are not publicly available, we split the training data to the training and validation sets following the split method in [18], which has also been widely adopted in other papers. Note that we train the models with the training set and evaluate the trained models with the validation set.

B. Training

We train our model with the KITTI dataset. We apply the following data augmentation strategies to the training data

- Flip: We apply random flip operations to all Lidar points along x axis.
- Translation: We apply random translation to all Lidar points by $-5 \sim 5$ meters along x and y axes and $-1 \sim 1$ meters along z axis.

Authorized licensed use limited to: Hanyang University. Downloaded on August 24,2023 at 05:00:46 UTC from IEEE Xplore. Restrictions apply.

TABLE I

DETECTION ACCURACY (AP) ON THE KITTI VALIDATION SET

Method	Modality	Easy	Mod.	Hard
MV3D [8]	RGB + Lidar	86.55	78.10	76.67
PC-CNN [15]	RGB + Lidar	83.61	77.36	69.61
F-PointNet [13]	RGB + Lidar	88.16	84.02	76.44
MV3D [8]	Lidar	86.18	77.32	76.33
PIXOR [9]	Lidar	86.79	80.75	76.60
Baseline	Lidar	87.35	78.49	76.97
Proposed method	Lidar	**88.78**	**84.64**	**78.57**

- Rotation: We apply random rotation to all Lidar points between $-5 \sim 5$ degrees around the Lidar position.
- Scaling: We randomly scale the Lidar points by the factor of $0.9 \sim 1.1$ along all three axes.

During training, we enable each augmentation strategy with probability of 0.5. This allows for enabling the multiple augmentation strategies at the same time. We train our model with the stochastic gradient descent (SGD) algorithm with the momentum parameter 0.9 and the initial learning rate 0.001. The learning rate decays by 0.1 at 193-th epoch and 257-th epoch. We set the weight decay to 0.0005 for L_2 regularization. During training, we use the group normalization method where the number of the groups is set to 2. Following [3], the focal loss parameters are set to $\alpha = 0.25$ and $\gamma = 2$. We do not use the pre-trained model for the whole network.

TABLE II

COMPARISON BETWEEN CONNET VS. SE METHOD

Method	Easy	Mod.	Hard
Baseline	87.35	78.49	76.97
SE method [19]	86.63	80.78	76.94
Our ConNet	**88.78**	**84.64**	**78.57**

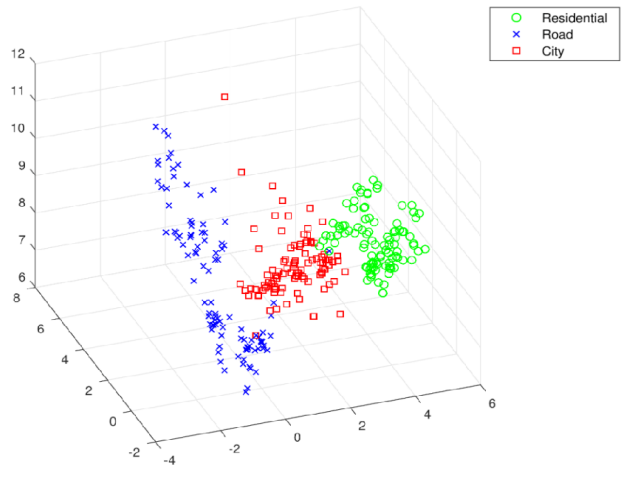


Fig. 2. Distribution of the context vectors for different traffic scenes

C. Experimental Results

In this section, we evaluate the detection accuracy of the proposed method. In order to demonstrate the benefit of the global context found by our ConNet, we compare our scheme with the baseline model which has the same structure but does not use the output of the ConNet. Note that the baseline model and proposed method are trained with the exactly same training procedure. Table I compares the average precision (AP) of the baseline and our object detector. We observe that our method outperforms the baseline detector by more than 2.5% for all difficulty levels. This result validates the effectiveness of the global context in performing BEV object detection. Note that the ConNet achieves up to 3.72% improvement over the baseline for the moderate difficulty category.

We also compare our algorithm with other state-of-the-art BEV object detectors (i.e., top rankers who have made their papers public in KITTI BEV object detection leaderboard). Table I shows that the proposed method achieves the best performance among the BEV object detectors using the Lidar sensor only. In addition, our BEV object detector is comparable to the methods using both Lidar and camera sensors.

Next, we investigate whether the structure of ConNet is indeed better than the CNN structure in capturing the geometric context from the 3D lidar points. As a reference, we replace ConNEt with the squeeze-and-excitation (SE) method [19], which extracts the contextual information from the 2D BEV representation. Table II shows that our ConNet achieves better performance than the SE-based method. This might be due to the fact that while the ConNet retains the 3D geometric structure of the environment in its input, the SE has lost some 3D information in the BEV representation.

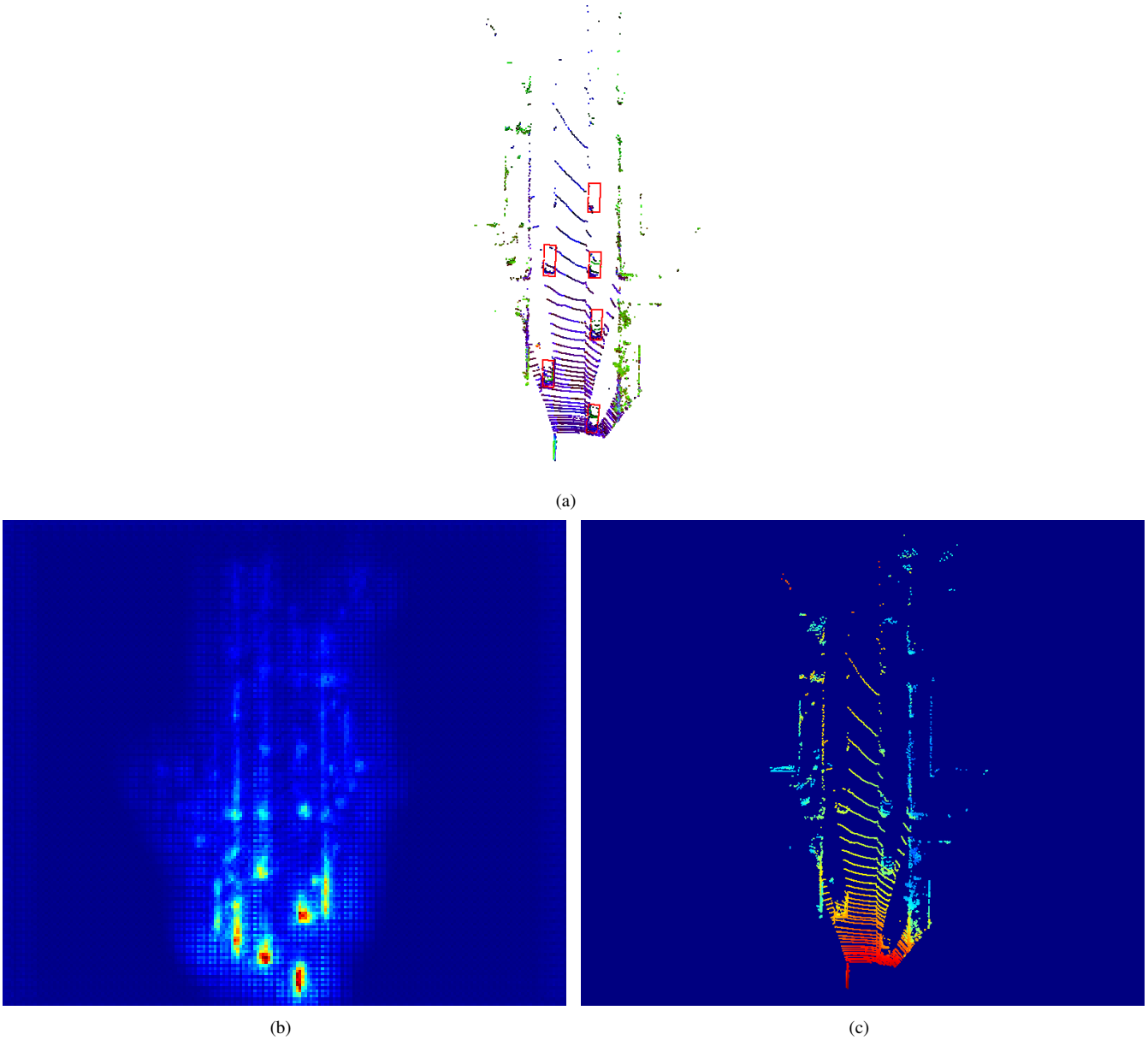
D. Behavior of Proposed Object Detector

In this section, we look into the behavior of the proposed network.

1) Distribution of Context Vector for Different Scene Categories: We take a look at the distribution of the context vector for the different scenes. The KITTI dataset has the examples collected from five different traffic scenarios. Among them, we randomly select 100 samples from three different scenarios; 1) city, 2) residential, and 3) road. For each test example, we extract the context vectors and reduce their dimensions into three using the principal component analysis (PCA). Fig. 2 shows how the resulting points are distributed in 3D space for different scene categories. We see that the points are clustered for the different traffic categories, which implies the context vector produced by our ConNet reflects the structure of the global scene in the background.

2) Visualization of Intermediate Features: In order to look at what part of features the DetNet and ConNet focus on, we visualize the intermediate features learned by both models. Fig. 3 (a) presents the projected BEV input image for the DetNet (for visualization, 36 channels are compressed into 3 channels). Fig. 3 (b) and (c) present the visualized features for the DetNet and ConNet at the position (1) and (2) in

Authorized licensed use limited to: Hanyang University. Downloaded on August 24,2023 at 05:00:46 UTC from IEEE Xplore. Restrictions apply.



(a)

(b)

(c)

Fig. 3. Visualization of the intermediate features produced by the DetNet and ConNet: (a) Input 2D BEV representation (objects in the red ground truth (GT) box), (b) the feature map from the DetNet, and (c) the feature map from ConNet

the Fig. 1, respectively. Note that the these feature maps for visualization is generated by using the magnitude of the 96-length feature vector as a pixel value. While the DetNet activates on the regions containing the objects, the ConNet rather focuses on the part of the road in the background. This shows that the DetNet and ConNet activate on different spatial regions to produce the good feature maps for BEV object detection.

V. Conclusions

In this paper, we proposed the enhanced BEV object detection method which uses the 3D global context inferred from the Lidar 3D points. We employ the ConNet to generate the $1 \times 1 \times k$ feature vector which captures the 3D geometrical structure around the surrounding. The context feature is concatenated with the CNN feature maps obtained from the main DetNet. When the combined feature maps are used for BEV object detection, our method achieves significant performance gain over the baseline detector. Our BEV object detector also outperforms the other detection methods using Lidar sensor only. Through some empirical analysis, we show that the proposed ConNet captures the structure on the global scene.

Authorized licensed use limited to: Hanyang University. Downloaded on August 24,2023 at 05:00:46 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: single shot multibox detector," in *Proc. European Conf. on Computer Vision (ECCV)*, 2016.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[3] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Inter. Conf. on Computer Vision (ICCV)*, 2017.

[4] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," 2017.

[5] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[6] B. Yang, M. Liang, and R. Urtasun, "Hdnet: Exploiting hd maps for 3d object detection," in *Conference on Robot Learning (CoRL)*, 2018.

[7] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[8] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[9] B. Yang, W. Luo, and R. Urtasun, "Pixor: Real-time 3d object detection from point clouds," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[10] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *Proc. IEEE/RSJ Inter. Conf. on Intel. Robots and Systems (IROS)*, 2018.

[11] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *Proc. European Conf. on Computer Vision (ECCV)*, 2018.

[12] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," *arXiv preprint arXiv:1812.04244*, 2018.

[13] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[14] K. Shin, Y. P. Kwon, and M. Tomizuka, "Roarnet: A robust 3d object detection based on region approximation refinement," *arXiv preprint arXiv:1811.03818*, 2018.

[15] X. Du, M. H. Ang Jr, S. Karaman, and D. Rus, "A general pipeline for 3d detection of vehicles," *arXiv preprint arXiv:1803.00387*, 2018.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] 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 *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Authorized licensed use limited to: Hanyang University. Downloaded on August 24,2023 at 05:00:46 UTC from IEEE Xplore. Restrictions apply.