FII-CenterNet: An Anchor-free Detector with Foreground Attention for Traffic Object Detection

Siqi Fan, Fenghua Zhu, Shichao Chen, Hui Zhang, Bin Tian, Yisheng Lv and Fei-Yue Wang

***Abstract*—Most successful object detectors are anchor-based, which is difficult to adapt to the diversity of traffic objects. In this paper, we propose a novel anchor-free method, called FII-CenterNet, which introduces the foreground information to eliminate the interference of the complex background information in traffic scenes. The foreground region proposal network segments the foreground based on boxes-induced segmentation annotation, and midground is proposed to provide rich edge information of the objects. In addition to foreground location, scale information is also introduced to improve the regression performance. Extensive experimental results on two public datasets verify the benefits of the introduction of the foreground information, and demonstrate that our FII-CenterNet achieves the state-of-the-art performance in both accuracy and efficiency.**

***Index Terms*—Object detection, Anchor-free detector, Foreground region proposal**

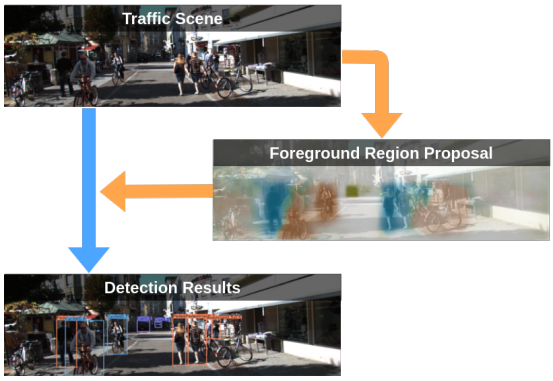

Fig. 1. Diagram of FII-CenterNet. Our method proposes the foreground region of traffic scenes based on semantic segmentation, and introduces foreground information to traffic object detection.

I. Introduction

INTELLIGENT transportation systems (ITS) are envisioned to bring great benefits to the development of smart cities and human societies. Object detection can locate traffic objects timely and accurately, and plays an increasingly important role in various ITS applications, such as autonomous driving. However, traffic object detection is also a challenging computer vision problem, which attracts the interests and efforts from both academia and industry.

The performance of object detection has been significantly improved in the past few years, with the successful application of deep convolutional neural networks (ConvNets) [1]. Most of the methods are anchor-based and can be divided into two categories, which are single-stage methods and two-stage methods respectively. Single-stage methods, such as SSD [2] and YOLO [3], complete the classification and regression of objects in one stage, and its advantage is that the detection speed is very fast. Two-stage methods, such as Fast R-CNN [4] and Faster R-CNN [5], can usually obtain more accurate detection results by introducing the region proposal network. Great efforts have been devoted to improve the accuracy and efficiency of anchor-based methods, and they are advancing toward maturity. However, these methods have some congenital deficiencies, which limit their popularity. Their detection performance relies heavily on the hyperparameters of the anchors, such as the size, aspect ratio and the numbers of anchors. As there is still no efficient method to adjust these hyperparameters autonomously, they have to be manually calibrated case by case.

To address this problem, anchor-free methods are put forward to improve the flexibility of the detectors [6] [7] and have been widely concerned. Anchor-free methods do not depend on preset anchors and can adapt to the diverse traffic objects through regression. However, their detection accuracy is ordinary, especially for the complex traffic scenes. The performance of traffic object detection can be further improved by eliminating the interference caused by the background information. In the anchor-based methods, the detection accuracy of the two-stage methods is better than that of the single-stage, which is largely due to the region proposal network (RPN). RPN distinguishes the foreground from the background. It can help to eliminate the interference of complex background

This work is supported in part by the National Key R&D Program of China 2018YFB1004803, NSFC U1811463, U1909204, 61773381, 61876011, Guandong Grant No.2019B1515120030, China Railway N2019G020, and the Intel Collaborative Research Institute for Intelligent and Automated Connected Vehicles (ICRI-IACV).(Corresponding author: Yisheng Lv)

Siqi Fan and Hui Zhang are with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China. Siqi Fan is also with the State Key Laboratory of Nuclear Power Safety Monitoring Technology and Equipment, China Nuclear Power Engineering Co., Ltd, Shenzhen of Guangdong Prov. 518172. (fansiqi2019@ia.ac.cn; zhanghui2015@ia.ac.cn)

Fenghua Zhu, Shichao Chen, Bin Tian, Yisheng Lv and Fei-Yue Wang are with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. Fenghua Zhu is also with Qingdao Academy of Intelligent Industries, Qingdao 266109, China. Shichao Chen is also with Faculty of Information Technology, Macau University of Science and Technology, Macau 999078, China. Fei-Yue Wang is also with Institute of Systems Engineering, Macau University of Science and Technology, Macau, China (fenghua.zhu@ia.ac.cn; shichao.chen@ia.ac.cn; bin.tian@ia.ac.cn; yisheng.lv@ia.ac.cn; feiyue.wang@ia.ac.cn)

information, and focus more on the features that are really useful for detection.

Inspired by the RPN in the anchor-based methods, we propose a novel anchor-free method by introducing foreground information to CenterNet, called **Foreground Information Introduction CenterNet (FII-CenterNet)**, which can achieve better detection accuracy with high efficiency. The diagram of our method is shown in Figure 1.

In sum, the main contribution of this paper are listed as follows:

- A foreground region proposal method based on semantic segmentation is proposed for anchor-free detectors.
- Midground is introduced as the transition region between foreground and background, which can provide rich edge information of the objects.
- Foreground scale information is introduced to regression process, which can improve the scale prediction performance of traffic objects.
- Our method is evaluated on two public datasets, KITTI and PASCAL VOC. Extensive experimental results demonstrate that FII-CenterNet effectively improves the traffic object detection performance and achieves the **state-of-the-art** performance in both accuracy and efficiency.

II. Related Work

Object detection has attracted wide attention for a long time, and great achievements and breakthroughs have been made in this field. Early object detection methods are based on hand-crafted features and classifiers, which is time-consuming. The representation abilities of hand-crafted features are very limited. CNN-based object detection methods can effectively extract features from images and perform end-to-end training. The existing CNN-based methods can be divided into two main categories according to whether they use anchor information, i.e., anchor-based detectors and anchor-free detectors.

A. Anchor-Based Detectors

These detectors are based on the preset anchor information, which can be further divided into two subcategories, i.e., single-stage detectors and two-stage detectors. The two-stage detectors have intermediate region proposal process, also known as region-based detectors. In contrast, single-stage detectors are also called region-free detectors.

1) Single-stage Detectors: These detectors [2] [3] [8] [9] [10] process classification and regression in one stage, which are efficient. SSD [2] and YOLO [3] are two typical single-stage detectors and they enable real-time object detection. YOLO V3 [8] utilizes feature pyramid network (FPN) and uses anchors with different sizes or aspect ratios to adapt various objects. DSSD [9] introduces additional large-scale context into object detection to improve the detection accuracy on small objects. Following, numerous researches have been carried out to optimize the detection accuracy by improving feature extraction networks and/or loss functions. Among them, RetinaNet [10] is one representative work, which uses FocalLoss to improve the imbalance of positive and negative samples in the image.

Although the detection speed of single-stage detectors is fast, the accuracy is average, while our FII-CenterNet performs well and maintains high detection efficiency.

2) Two-stage Detectors: These detectors [11] [4] [5] are usually composed of region proposal generation part and detection part. R-CNN [11] performs region proposal through selective search. Fast-RCNN [4] effectively improves computational efficiency by extracting features from the full image and crops features instead. Faster R-CNN [5] merges RPN with Fast-RCNN into a unified detection network by sharing feature maps. RPN is proposed to generate high-quality region proposals, which samples fixed-shape anchors and classifies each into foreground or background.

The region proposal process effectively distinguishes the foreground and background regions, and eliminates the interference of complex background information. However, it also increases the computational complexity of detection.

The main deficiency of anchor-based detectors is that the anchors must be set manually. Usually, the tuning processes of these anchors are time-consuming and difficult, which limits the usability of these methods, especially for complex traffic scenes with diverse objects.

B. Anchor-Free Detectors

These detectors do not depend on the pre-set anchors by using keypoints estimation for object detection, which can be roughly divided into two subcategories according to the number of keypoints, i.e., methods based on the joint expression of multiple keypoints and methods based on a single center point per object.

1) Object Detection by Multiple Keypoints Estimation: These methods [7] [12] [13] [14] [15] are based on the joint expression of multiple keypoints. CornerNet [7] detects two bounding box corners as keypoints, which is sensitive to the edges and causes false detection. To improve the accuracy, Keypoint Triplets for Object Detection [13] introduces geometric constraints. It detects two corners and the center points, and then verifies whether there is a center point in the center area determined by the corners. ExtremeNet [14] decomposes corner detection into detecting the four-sided extreme points of the bounding box. The detection result is expressed in a more accurate way, instead of the rectangle bounding box. RepPoints [15] uses 9 keypoints to get a more detailed description, and uses weak supervision to locate them.

Generally, object detection by multiple keypoints estimation can effectively improve the accuracy. However, they require a combinatorial grouping stage after keypoint detection, which significantly slows down the detection speed. Introducing foreground information can improve the detection accuracy with less time-consuming increment.

2) Object Detection by Center Point Estimation: These methods [16] [6] [17] [18] simply extract a single center point per object without the need for grouping or post-processing. DenseBox [16] uses FCN (Fully Convolution Network) for prediction. It estimates the probability of a pixel as a center

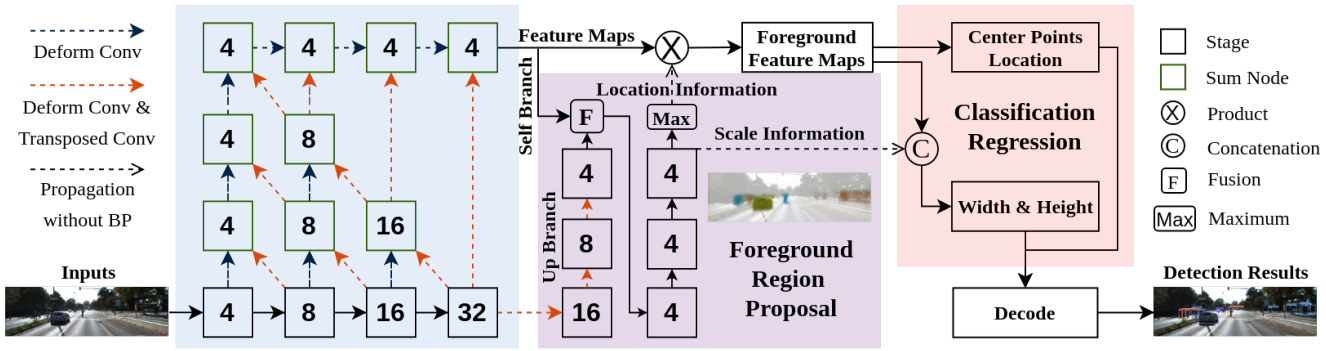

Fig. 2. FII-CenterNet: introducing foreground information to CenterNet for traffic object detection. The numbers in the boxes represent the stride to the image. We use modified DLA-34 in CenterNet to extract features. It uses deformable convolution to change the channels and uses transposed convolution to upsample the feature maps. We show deformable convolution as a blue dashed arrow and show these two steps together as an orange dashed arrow. The black dashed arrow means propagation without backward propagation.

point, and estimates the relative displacement of the two corners of the bounding box centered on it. CenterNet [6] simplifies object detection into center point detection and scale prediction. CSP [17] is similar to it, which generates center point heatmap and scale prediction map based on the extracted feature map. FCOS [18] is based on FPN, which performs object detection on multi-scale feature maps. Besides, it also proposes center-ness to assist training.

Object detection by center point estimation is composed of center points detection and scale prediction, which can effectively adapt to the traffic object with variable size. In addition, it can meet the needs of detection efficiency in traffic scenarios. However, the detection accuracy in complex traffic scenes is ordinary. To deal with that, we introduce foreground region proposal to eliminate the interference of complex background information and improve the accuracy of traffic object detection.

C. Detectors Exploiting Segmentation Information

Semantic segmentation is another important computer vision task and is widely used in ITS, such as road detection [19]. Our foreground region proposal method is based on that. We are not the first one to show segmentation information can help object detection. Mask R-CNN [20] shows that multi-task training can help to improve the object detection task. He et al.[20] and Shrivastava et al. [21] trained the object detection task with instance segmentation annotation. Our work only uses bounding box annotation and does not consider extra annotation. Gidaris and Komodakis [22] concatenated semantic segmentation features with detection features at the highest level, while DES [23] uses activation instead of concatenation and combines the two features at the lowest detection feature map. However, we use semantic segmentation features to get foreground region information. The location information is used to generate foreground region feature map, and the scale information is introduced to regression task. In addition, their works are based on anchor-based detectors. Gidaris and Komodakis [22] utilize Faster R-CNN and DES [23] is based on SSD. Our work is on the basis of CenterNet, which is a typical anchor-free detector.

III. FII-CenterNet Approach

The proposed FII-CenterNet is an anchor-free detection network. Built on CenterNet, a foreground region proposal network is added to introduce the foreground information. The structure of the FII-CenterNet is shown in Figure 2.

FII-CenterNet uses a modified DLA-34 in CenterNet to extract features. Deep Layer Aggregation (DLA) [24] is an image classification network with hierarchical skip connections. The modified DLA-34 utilizes deformable convolution as skip connection from lower layers to the output. Specifically, the original convolution is replaced with 3×3 deformable convolution [25] at every upsampling layer.

Foreground region proposal network aims to estimate the foreground region. By introducing the foreground location information, foreground feature maps are generated from the feature maps extracted by DLA-34. Mathematically, let $\mathbf{FM}$ be the feature map and $\mathbf{FFM}$ be the foreground feature map.

$$\mathbf{FFM} = \mathbf{FM} \odot \mathbf{F}$$

where $\mathbf{F}$ is the foreground region proposal result, and $\odot$ is the element-wise multiplication.

For an input image of width W and height H, a center point heatmap $\mathbf{P} \in [0,1]^{\frac{W}{R} \times \frac{H}{R} \times C}$ is produced in classification process, where R is the output stride and C is the object categories in traffic object detection. Thus, the prediction $P_{xyc} = 1$ corresponds to a detected center point. The ground truth center point heatmap $\mathbf{G_p} \in [0,1]^{\frac{W}{R} \times \frac{H}{R} \times C}$ is generated using a Gaussian kernel $G_{p_{xyc}} = exp(-\frac{(x-\hat{cp}_x)^2+(y-\hat{cp}_y)^2}{2\sigma^2_{\hat{cp}}})$, where $\hat{cp} = \frac{cp}{R}$ is the low-resolution equivalent point of the ground truth center point $cp \in R$ of class c, and $\sigma_{\hat{cp}}$ is an object size-adaptive standard deviation [7]. At inference time, the top 100 peaks in the heatmap whose value is not less than its 8-connected neighbors are kept as center points prediction.

A scale prediction map $\mathbf{S} \in \mathbb{R}^{\frac{W}{R} \times \frac{H}{R} \times 2}$ is produced in regression process. For center point p_k, the scale prediction is $S_k = (w_k, h_k)$, where w_k and h_k correspond to the width and height of the object centered at p_k.

The local offset $\mathbf{O} \in \mathbb{R}^{\frac{W}{R} \times \frac{H}{R} \times 2}$ is also predicted additionally as that in CenterNet. All classes share the same offset prediction which is to recover the discretization error caused

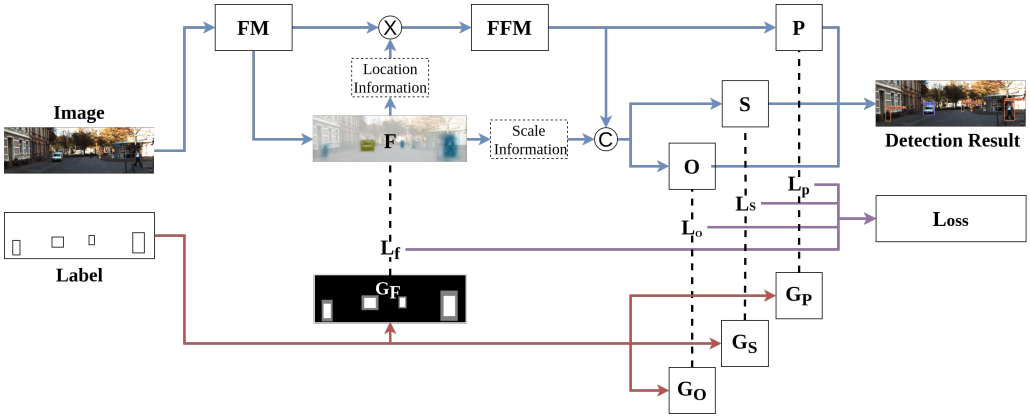

Fig. 3. FII-CenterNet training diagram. The loss function can be divided into four parts. L_f is the loss for foreground region proposal network; L_p is the loss for center points prediction network; L_s is the loss for scale prediction network; L_o is the loss for offset prediction network.

by the output stride. For center point p_k, the offset prediction is $O_k = (\delta_{k_x}, \delta_{k_y})$.

The final detection results are described as bounding box. For the center point p_k, the corresponding bounding box is $(x_{k_1}, y_{k_1}, x_{k_2}, y_{k_2})$, which is decoded with p_k and S_k. Specifically,

$$x_{k_1} = p_{k_x} + \delta_{k_x} - \frac{w_k}{2}$$

$$y_{k_1} = p_{k_y} + \delta_{k_y} - \frac{h_k}{2}$$

$$x_{k_2} = p_{k_x} + \delta_{k_x} + \frac{w_k}{2}$$

$$y_{k_2} = p_{k_y} + \delta_{k_y} + \frac{h_k}{2}$$

The following subsections are organized as follows. First of all, we propose the foreground region proposal network in section A. Secondly, we describe the way to train the network. Loss function and foreground segmentation label generation method are described in section B and C, respectively. Some edge information of the objects are lost because of the box-induced label. Thus, we propose the concept of midground in section D. In section E, we introduce the scale information to regression network to make full use of the foreground information. Training diagram is shown in Figure 3.

A. Foreground Region Proposal Network

In order to propose foreground region based on semantic segmentation, there are two feasible implementation methods, foreground region proposal by up branch and foreground region proposal by self branch.

1) Foreground Region Proposal by Up Branch: In this method, foreground region proposal is on the basis of the results of the encoder, as shown in the up branch in Figure 2. It performs an upsampling operation through an additional branch and proposes the foreground region using the results obtained by multiple consecutive convolutions. In the up branch, deformable convolution is used to change the channels, and transposed convolution is used to upsample the feature map. The encoder-decoder structure is commonly used in semantic segmentation networks.

Mathematically, let $\mathbf{E}$ be the encoded features, this method computes foreground region proposal $\mathbf{F}$ as

$$\mathbf{F} = \mathcal{F}(\mathcal{D}(\mathbf{E}))$$

where $\mathcal{D}(\mathbf{E})$ is the intermediate result decoded from $\mathbf{E}$ using deformable convolution and transposed convolution.

2) Foreground Region Proposal by Self Branch: This method directly performs further convolution operations on the basis of feature maps, as shown in self branch in Figure 2. It directly uses the extracted feature maps and does further analysis and processing. The foreground region proposal is obtained by multiple consecutive convolutions.

$$\mathbf{F} = \mathcal{F}(\mathbf{FM})$$

Theoretically, the above two methods are both feasible. The final choice is depend on the detection performance, so the two methods are integrated in the network. The results of self branch and up branch are fused, and then three consecutive convolutions are performed to propose the foreground region. We will describe the fusion methods of the two branches and the final choice later in section IV.B.

To obtain better proposals, the foreground region is estimated under different categories first. The prediction $F_{xyc} = 1$ corresponds to a proposed foreground pixel. Then the pixel-wise maximum value of the proposals under different categories is calculated to get the final foreground region proposal.

B. Loss Function

The loss function can be divided into four parts:
(a) L_f is the loss for foreground region proposal network;
(b) L_p is the loss for center points prediction network;
(c) L_s is the loss for scale prediction network;
(d) L_o is the loss for offset prediction network.
For L_f and L_p, a modified focal loss [7] is used.

$$L_f = \frac{-1}{N} \sum_{xyc} \begin{cases} (1 - F_{xyc})^{\alpha} log(F_{xyc}) & G_{f_{xyc}} = 1 \\ (1 - G_{f_{xyc}})^{\beta} (F_{xyc})^{\alpha} & \\ log(1 - F_{xyc}) & otherwise \end{cases}$$

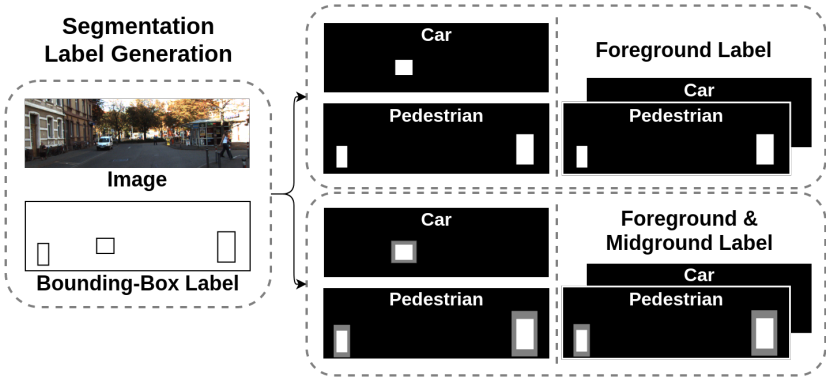

Fig. 4. Schematic diagram of segmentation label generation: The segmentation labels are generated from the bounding box labels. The top right is the generation of foreground labels. Midground is introduced to labels as shown in bottom right.

where F_{xyc} is foreground pixels proposal, $G_{f_{xyc}}$ is the groundtruth of that, α and β are the hyper-parameters, and N is the normalization factor. We use $\alpha = 2$ and $\beta = 4$ following Law and Deng [7], and N is the number of pixels where $G_{f_{xyc}} = 1$.

$$L_p = \frac{-1}{N} \sum_{xyc} \begin{cases} (1 - P_{xyc})^{\alpha} log(P_{xyc}) & G_{p_{xyc}} = 1 \\ (1 - G_{p_{xyc}})^{\beta} (P_{xyc})^{\alpha} & \\ log(1 - P_{xyc}) & otherwise \end{cases}$$

where P_{xyc} is the prediction of center points, and $G_{p_{xyc}}$ is the groundtruth of that. We also set $\alpha = 2$ and $\beta = 4$.

For L_s and L_o, an L_1 loss is used.

$$L_s = \frac{1}{N} \sum_{k=1}^{N} |S_k - G_{s_k}|$$

where S_k is the scale prediction at center point p_k, and G_{s_k} is the groundtruth of that.

$$L_o = \frac{1}{N} \sum_{k=1}^{N} |O_k - G_{o_k}|$$

where O_k is the offset prediction at center point p_k, and $G_{o_k} = \frac{cp}{R} - \hat{cp}$ is the corresponding groundtruth.

The overall loss function is

$$L = \lambda_f L_f + \lambda_p L_p + \lambda_s L_s + \lambda_o L_o$$

where $\lambda_f, \lambda_p, \lambda_s$ and λ_o are the loss weights corresponding to the four parts.

C. Foreground Segmentation Label Generation

The foreground segmentation label is generated to train the foreground region proposal network, which is a kind of boxes-induced segmentation annotation [26].

We first project the groundtruth bounding box into the corresponding location under the output stride. The pixel in the label will be set to 1, if it locates within the projected bounding box, otherwise set to 0. The size of label is $\frac{H}{R} \times \frac{W}{R} \times C$, where H and W are the height and width of the image, R is the output stride, and C is the object categories in traffic object detection.

D. Midground

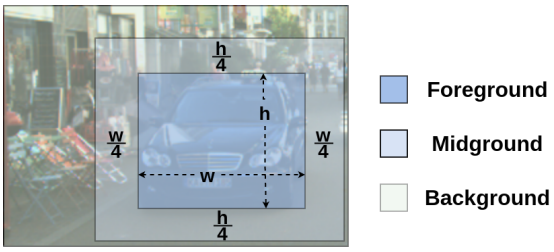

Fig. 5. Schematic diagram of the midground. Midground is the ring-shaped area outside the foreground.

It is worth noting that some areas in the two-stage detectors' RPN are classified as neither foreground nor background. An anchor in Faster R-CNN [5] is labeled based on the overlap with any groundtruth object.

$$FasterR-CNN[5] \begin{cases} foreground & overlap > 0.7 \\ background & overlap < 0.3 \\ ignored & otherwise \end{cases}$$

Those ignored areas are the transition zone. In order to locate the object accurately, the bounding box is usually close to the edge of the object, which causes the resulting segmentation label generated on the basis of that is a strict foreground region, rarely including the surrounding area of the object. The surrounding areas of the objects contain the edge information, which is important to both the classification and the regression. To deal with that, the concept of midground is introduced. Midground region is defined as the ring-shaped area outside the foreground region, as shown in Figure 5.

The pixels in midground region are set to 0.5 while generating the segmentation label. The factor $(1 - G_{f_{xyc}})^{\beta}$ in L_f makes the loss function give a higher tolerance to the midground region. To a certain extent, the transition between foreground and background area is realized by midground.

E. The introduction of Foreground Scale Information

The location information is only part of the foreground information. It is worth noting that, due to the foreground

TABLE I
STRUCTURE EXPERIMENTS ON KITTI VALIDATION SET

	Car			Pedestrian			Cyclist			Runtime
	Easy(%)	Moderate(%)	Hard(%)	Easy(%)	Moderate(%)	Hard(%)	Easy(%)	Moderate(%)	Hard(%)	/ms
CenterNet	91.6 ± 0.9	87.5 ± 0.2	79.1 ± 0.2	75.5 ± 0.3	65.9 ± 0.3	57.5 ± 0.5	78.1 ± 1.0	58.3 ± 0.7	55.6 ± 0.6	69
Up Branch	92.0 ± 0.9	88.4 ± 0.1	79.7 ± 0.1	75.7 ± 0.4	67.6 ± 0.6	59.1 ± 0.6	77.8 ± 0.7	57.8 ± 0.7	55.2 ± 0.6	80
Self Branch	**94.0 ± 1.3**	**88.7 ± 0.1**	**79.8 ± 0.1**	**77.8 ± 0.1**	**68.6 ± 0.6**	**61.0 ± 0.4**	80.3 ± 1.0	58.7 ± 0.7	55.6 ± 0.6	79
Fusion(Summation)	93.8 ± 1.4	88.4 ± 0.1	79.6 ± 0.1	77.4 ± 0.3	67.6 ± 0.3	60.5 ± 0.2	80.1 ± 0.4	59.8 ± 0.4	56.9 ± 0.7	84
Fusion(Concatenation)	92.6 ± 1.1	88.5 ± 0.1	79.7 ± 0.1	76.2 ± 0.3	66.5 ± 0.3	58.2 ± 0.7	**81.2 ± 1.2**	**60.1 ± 1.0**	**57.3 ± 0.9**	87

region proposal network is category-based, the proposals of each category actually contain the scale information of the object. Introducing the foreground scale information into the regression process can assist the scale prediction. Therefore, the foreground region proposals under different categories are concatenated with the foreground feature maps in the FII-CenterNet as the input of regression.

IV. EXPERIMENTS

In this section, we evaluate the performance of the proposed FII-CenterNet on two public datasets, KITTI [27] and PASCAL VOC [28]. The experiments are implemented in the Pytorch on a machine with NVIDIA Titan Xp GPUs, CUDA 9.0 and cuDNN v7.

A. Implementation Details

1) Experiments on KITTI: The experimental dataset is obtained from the KITTI object detection benchmark, which consists of 7481 labeled images and 7518 testing images. For the KITTI benchmark does not provide groundtruth labels for the testing set, we follow standard training and validation splits in literature [29] [30] to implement our structure experiments and ablation experiments. The labeled images are divided into training set with 3712 images and validation sets with 3769 images. For cars it requires an overlap of 70%, while for pedestrians and cyclists it requires an overlap of 50% for a detection. Detections in 'don't care' areas or detections which are smaller than the minimum size do not count as false positive. Three-level difficulties including 'Easy', 'Moderate' and 'Hard' are defined in literature [27].

The original image resolution is kept and the size is pad to 1280×384. We set all of the loss weight to 1 in experiments on KITTI. The network is trained on one GPU with the batch size of 8. We use Adam optimizer [31] with a initial learning rate of 10^{-4}. It is trained for 70 epochs, with learning rate dropped at the 45 (10^{-5}) and 60 (10^{-6}) epoch, respectively. Flip augmentation is used in testing.

2) Experiments on PASCAL VOC: The dataset is collected by Everingham et al. [28] for visual object category recognition and detection. We train on VOC 2007 and VOC 2012 trainval sets, and test on VOC 2007 test set. It contains 16551 training images and 4962 testing images. The annotated images are compiled from the Flickr photo-sharing website and has great variability in object size, orientation pose, illumination position and occlusion. The original dataset consists of 20 categories. We only evaluate on the four concerned traffic object categories: 'car', 'bicycle', 'motorbike' and 'person'. The evaluation metric is mean average precision (mAP) of the four traffic object categories.

We experiment our FII-CenterNet in a small training resolution. The input size is 384×384. We set $\lambda_s = 0.1$, while all other hyper-parameters in loss function are the same as the KITTI experiments. The network is trained on one GPU with the batch size of 32. Adam optimizer [31] is used with an initial learning rate of 1.25×10^{-4}. It is trained for 140 epochs, with learning rate dropped at the 90 (1.25×10^{-5}) and 120 (1.25×10^{-6}) epoch, respectively. Flip augmentation is used in testing.

B. Structure Experiments on KITTI Validation Set

We evaluate the structure of the foreground region proposal network in this subsection. Two feasible foreground region proposal methods are proposed in the previous section, which correspond to the up branch and the self branch in the network. We conduct experiments on the two structures, and explore the performance under different fusion methods at the same time. Two common fusion methods are used, i.e., summation and concatenation. The results are reported in Table I.

The validation AP fluctuates by up to 10% for the small recall thresholds. We thus train 5 models per experiment and report the average with standard deviation. The baseline for all experiments in this subsection is the original CenterNet.

First of all, no matter which foreground region proposal method is adopted, almost all of the performances are better than the baseline. The improvements demonstrate the effectiveness of the introduction of foreground location information.

Secondly, when evaluating our two structures and the two fusion methods of them, it can be seen that the self branch has the best performance for car and pedestrian detection. For moderate difficulty, the detection accuracy of car and pedestrian are improved by **1.2%** and **2.7%**, respectively. There is a **3.5%** enhancement for pedestrian's hard AP (an indicator influenced largely by small object detection), indicating that the structure helps to eliminate the interference of complex background information. For the detection of cyclist, the performance of the concatenation is the best, and the detection accuracy of moderate difficulty is increased by **1.8%**. Besides, it can be seen that the two fusion methods have different performances on the three categories. Concatenation could maintain complete information outputted from previous operations/layers, while summation simply sums the outputs so that it generally loses part of the information. Moreover, summation can be regarded as a special case of concatenation for the multichannel convolution operation, in which

Fig. 6. Visualization examples of FII-CenterNet for traffic object detection. Left: the result of foreground region proposal; Middle: the detected center points; Right: the final detection result.

the convolutional weights are same for all the channels and the weights between the channels are set to one manually. Therefore, the model using concatenation is more complex and is more suitable for the detection of complex object, but it may lead to overfitting for the simple object detection at the same time. Hence, concatenation achieves a better performance for detecting the most complex object 'cyclist' among the three testing objects, while its performance is close or even poorer than summation for detecting cars and pedestrians.

It can be seen that while the structures effectively improve the performance of traffic object detection, they also slow down the detection speed. Different methods slow down the speed to different extents, among which the self branch introduces the least time-consuming increment.

Theoretically, foreground region proposal by the self branch is a better choice. The reasons are listed as follow.

Analysis from the computational complexity. Foreground region proposal by the self branch is directly based on the results of feature maps, no additional decode operation is required. Thus the additional computational complexity is relatively small. Correspondingly, it will consume less time, and is more suitable for the field of intelligent transportation systems with high requirement of detection efficiency.

Analysis from the detection network structure. It produces the foreground region proposal on the basis of feature maps, which makes the feature extraction network pay more attention to the foreground region to get more accurate proposals. Therefore, the feature maps will contain more features about the foreground region, which can help to improve the performance of both classification and regression.

The experimental results meet the expectations of the theoretical analysis above. Taking into account that the occurrence frequency of car and pedestrian is higher than that of cyclist in actual traffic scenarios, and considering the object detection efficiency at the same time, the self branch is finally selected as the actual structure. The following experiments are also carried out on the basis of this structure.

C. Ablation Experiments on KITTI Validation Set

The baseline for all experiments in this subsection is the basic FII-CenterNet, which only introduces the location information of the foreground region. We explore the effectiveness of the midground and the introduction of foreground scale information in this subsection. For the same reason, five models are trained per experiment and the average with standard deviation is reported. The results are shown in Table II.

TABLE II
ABLATION EXPERIMENTS ON KITTI VALIDATION SET

	Car			Pedestrian			Cyclist		
	Easy(%)	**Moderate(%)**	**Hard(%)**	**Easy(%)**	**Moderate(%)**	**Hard(%)**	**Easy(%)**	**Moderate(%)**	**Hard(%)**
baseline	94.0 ± 1.3	88.7 ± 0.1	79.8 ± 0.1	77.8 ± 0.1	68.6 ± 0.6	61.0 ± 0.4	80.3 ± 1.0	58.7 ± 0.7	55.6 ± 0.6
+ scale information	93.3 ± 0.9	88.7 ± 0.1	79.8 ± 0.1	77.6 ± 0.7	69.7 ± 0.5	61.2 ± 0.4	**81.4 ± 1.0**	**59.7 ± 0.4**	**56.9 ± 0.4**
+ midground	**94.9 ± 1.0**	**88.8 ± 0.1**	**80.0 ± 0.2**	77.3 ± 0.4	68.6 ± 0.6	60.1 ± 0.7	79.4 ± 0.8	59.1 ± 0.5	56.2 ± 0.6
+ scale information & midground	93.3 ± 0.8	**88.8 ± 0.1**	**80.0 ± 0.1**	**78.5 ± 0.6**	**70.1 ± 0.3**	**61.3 ± 0.3**	79.7 ± 0.7	59.6 ± 0.7	56.5 ± 0.6

TABLE III
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON KITTI TEST SET

		Car			Pedestrian			Cyclist		
	FPS	**Easy(%)**	**Moderate(%)**	**Hard(%)**	**Easy(%)**	**Moderate(%)**	**Hard(%)**	**Easy(%)**	**Moderate(%)**	**Hard(%)**
TuSimple [32] [33]	0.63	95.12	94.47	86.45	88.87	78.40	73.66	83.68	75.22	65.22
MonoPair [34]	16.67	96.61	93.55	83.55	78.81	61.57	56.51	74.77	56.37	48.37
RRC [35]	0.28	95.68	93.40	87.37	85.98	76.61	71.47	86.81	76.81	66.59
sensekitti [36]	0.22	94.79	93.17	84.38	82.72	68.41	62.72	82.90	73.48	64.03
SJTU-HW [37] [38]	1.18	96.30	93.11	82.21	87.17	75.81	69.86		/	
EAS [39]	3.70	93.91	91.02	77.93	86.71	76.07	70.02		/	
Deep3DBox [40]	0.67	94.71	90.19	76.82		/		84.36	74.78	64.05
SubCNN [41]	0.50	94.26	89.98	79.78	84.88	72.27	66.82	79.36	71.72	62.74
3DOP [42]	0.33	92.96	89.55	79.38	83.17	69.57	63.48	80.52	68.71	61.07
Mono3D [43]	0.24	94.52	89.37	79.15	80.30	67.29	62.23	77.19	65.15	57.88
MS-CNN [44]	2.50	93.87	88.68	76.11	85.71	74.89	68.99	84.88	75.30	65.27
Faster R-CNN [5]	0.50	88.97	83.16	72.62	79.97	66.24	61.09	72.40	62.86	54.97
Ours	11.11	94.48	91.03	83.00	81.32	67.31	61.29	79.04	66.54	57.76

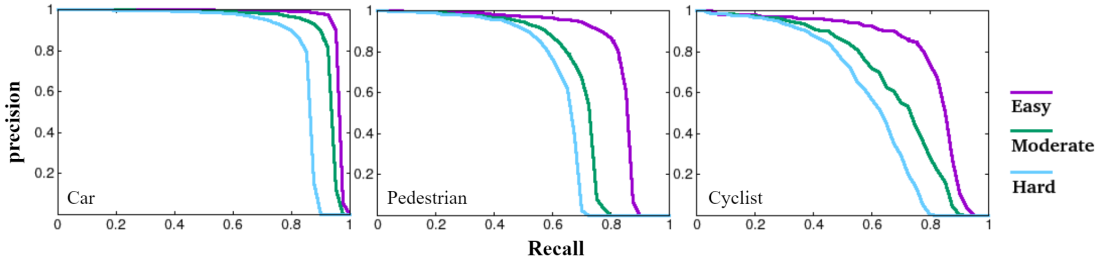

Fig. 7. Precision-recall curves on KITTI benchmark. From left to right are the precision-recall Curve of car, pedestrian and cyclist. For moderate difficulty, the detection accuracy of the three typical traffic object categories achieve 91.03%(car), 67.31%(pedestrian) and 66.54%(cyclist), respectively.

First of all, the improvement from the first row to the second row in the table demonstrates that the introduction of scale information can effectively improve the detection performance. The improvement for pedestrian and cyclist detection is more obvious. For Moderate difficulty, the detection accuracy of pedestrian and cyclist has been improved by **1.1%** and **1.0%**, respectively. As a matter of fact, the detection performance of car has also been slightly improved, but it cannot be seen from the table due to rounding.

Secondly, the effectiveness of the midground is verified in the third row. It can be seen that the introduction of the midground makes the detection accuracy of car and cyclist improved. For moderate difficulty, the accuracy is improved by **0.1%** and **0.4%**, respectively. It is increased by **0.2%** and **0.6%** for hard AP. But at the same time, the detection performance of pedestrian has declined to a certain extent.

Finally, when both the foreground scale information and midground are introduced, the detection accuracy of car and pedestrian reaches the best. The detection accuracy is improved by **0.1%** and **1.5%** for moderate AP, respectively, and it is **0.2%** and **0.3%** for hard AP. However, for cyclist detection, the accuracy is best when only the foreground scale information is introduced. The introduction of midground has a certain effect on improving the performance of traffic object detection, but sometimes it will reduce the promotion effect of the introduction of foreground scale information. Considering the transition effect of midground, we think that it may introduce uncertainty into the scale information, which ultimately leads to the weakening phenomenon.

The visualization examples of FII-CenterNet on KITTI validation set are shown in Figure 6.

D. Comparisons with Other Approaches on KITTI Test Set

In order to compare with other state-of-the-art approaches, we trained our FII-CenterNet with 7481 labeled data, and then submitted the results to the KITTI leaderboards. Our FII-CenterNet achieves the **state-of-the-art** performance in both accuracy and efficiency on KITTI benchmark.

For moderate difficulty, the detection accuracy of the three typical traffic object categories achieve **91.03% (car)**, **67.31% (pedestrian)** and **66.54% (cyclist)** with high detection efficiency. The precision-recall curves are shown in Figure 7. The

TABLE IV
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON PASCAL VOC 2007 TEST SET

Methods	Backbone	Input Size	mAP(%)	FPS	Bicycle(%)	Car(%)	Motorbike(%)	Person(%)
Single-Stage Detectors								
RON384 [45]	VGG-16	384×384	81.3	15.0	82.4	84.3	82.4	76.2
SSD300 [2]	VGG-16	300×300	80.8	58.0	80.2	84.2	82.6	76.2
SSD300+VSSA(Horizontal) [46]	MobileNet	300×300	84.0	/	88.3	86.7	84.6	76.4
SSD300+VSSA(Vertical) [46]	MobileNet	300×300	83.8	/	88.4	87.5	83.4	75.8
SSD512 [2]	VGG-16	512×512	84.0	23.0	84.7	87.5	83.9	79.7
SSD321 [9]	ResNet-101	321×321	83.3	11.2	84.6	84.0	85.4	79.1
SSD513 [9]	ResNet-101	513×513	86.5	6.8	87.5	88.1	87.5	83.0
DSSD321 [9]	ResNet-101	321×321	84.4	9.5	84.9	86.2	86.7	79.7
DSSD513 [9]	ResNet-101	513×513	86.5	5.5	86.2	88.7	87.5	83.7
Two-Stage Detectors								
Fast R-CNN [4]	VGG-16	$\approx 1000 \times 600$	75.8	0.5	78.1	78.6	76.6	69.9
Faster R-CNN [5]	VGG-16	$\approx 1000 \times 600$	79.5	7.0	79.0	84.7	77.5	76.7
Faster R-CNN [5]	ResNet-101	$\approx 1000 \times 600$	81.3	2.4	80.7	85.3	80.9	78.4
ION [47]	VGG-16	$\approx 1000 \times 600$	80.0	1.3	79.2	84.2	81.3	75.3
R-FCN [48]	ResNet-101	$\approx 1000 \times 600$	84.2	9.0	87.2	88.5	79.9	81.2
Detectors Exploiting Segmentation Information								
MR-CNN [22]	VGG-16	$\approx 1000 \times 600$	82.9	0.03	84.1	85.9	85.0	76.4
DES [23]	VGG-16	300×300	85.4	/	86.0	87.3	87.5	80.8
Shrivastava et al. [21]	VGG-16	$\approx 1000 \times 600$	81.7	/	80.5	86.5	81.6	78.2
Foreground			**FII-CenterNet**					
Loaction **Scale** **Midground**	**Backbone**	**Input Size**	**mAP(%)**	**FPS**	**Bicycle(%)**	**Car(%)**	**Motorbike(%)**	**Person(%)**
	DLA-34	384×384	83.7	35.1	84.8	84.9	84.2	80.7
✓	DLA-34	384×384	84.9		86.4	85.9	86.0	81.4
✓ ✓	DLA-34	384×384	85.6		86.7	86.5	87.0	82.1
✓ ✓ ✓	DLA-34	384×384	86.2	30.0	87.3	87.1	87.8	82.7

pedestrian detection accuracy can be improved to 71.01%, if multi-scale augmentation is used in testing.

Table III makes the comparison between the FII-CenterNet and other published vision-based methods on KITTI benchmark. We only compare with the method which is evaluated on at least two traffic object categories. The comparison is mainly based on the moderate difficulty of the three-level difficulties. For detection accuracy, only TuSimple[32] [33], RRC[35] and sensekitti[36] are better than our method on all the three traffic categories. The detection efficiency of FII-CenterNet is better than most of the methods, except for MonoPair [34]. However, the detection accuracy on pedestrian and cyclist of MonoPair [34] are 61.57% (5.74% lower) and 56.37% (10.17% lower), respectively.

The typical two-stage method, Faster R-CNN [5], achieves 83.6%, 66.24% and 62.86% on the three categories. FII-CenterNet performs better than it with a large margin, not to mention the detection speed. Our method also has a comparable accuracy of car detection with EAS [39], which is a recent state-of-the-art anchor-based detector for traffic object detection. FII-CenterNet also achieves comparable performance with other prevalent detectors, i.e., Deep3DBox [40], 3DOP [42] and Mono3D [43], at a faster speed. Deep3DBox is based on slow-RCNN, 3DOP is on the basis of Fast-RCNN, and Mono3D is Faster R-CNN based.

E. Experiments on PASCAL VOC 2007

We further compare our FII-CenterNet with other representative methods on PASCAL VOC 2007, which is reported in Table IV. All methods are trained on VOC 2007 and VOC 2012 trainval sets and tested on VOC 2007 test set. Our FII-CenterNet also achieves the **state-of-the-art** performance in both accuracy and efficiency on VOC 2007.

FII-CenterNet achieves **86.2%** mAP for traffic object detection when using input size of 384×384. For the four concerned traffic object categories, FII-CenterNet achieves **87.3% (bicycle), 87.1% (car), 87.8% (motorbike) and 82.7% (person)**, respectively. The corresponding precision-recall curves are shown in Figure 8 (a). Besides, several versions of FII-CenterNet are evaluated on VOC 2007 test set. The ablation experiments results verify the effectiveness of each module.

First of all, FII-CenterNet outperforms all single-stage methods using such small input size, e.g., RON384 [45], SSD300 [2], DSSD321 [9]. The first section in Table IV. contains some representative single-stage methods. Among the methods using small input size, DSSD321 [9] has the best detection accuracy. It achieves 84.4% mAP, which is still **1.8%** lower than our method. Furthermore, these methods can be improved by using the larger input size. DSSD improves the mAP from 84.4% to 86.5% due to the input size, but the detection efficiency is decreased. Maintaining the small input size, our method has a comparable performance.

Secondly, we also compare our method with some representative two-stage detectors, e.g., Fast R-CNN [4], Faster R-CNN [5], ION [47] and R-FCN [48], which is shown in second section of Table IV. There are two different version of Faster R-CNN. Faster R-CNN with VGG-16 backbone network achieves 79.5% mAP, while Faster R-CNN with ResNet-101 backbone network achieves 81.3% mAP. Our method performs better than it with a large margin. R-FCN [48] has the best detection accuracy among the two-stage methods shown in Table IV. When compared with it, FII-

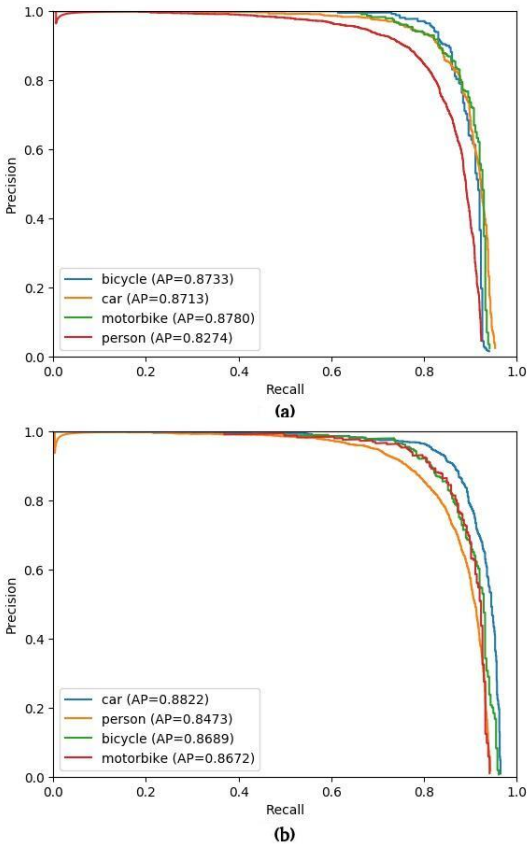

Fig. 8. Precision-recall curves on PASCAL VOC 2007 test set. All methods are trained on VOC 2007 and VOC 2012 trainval sets and tested on VOC 2007 test set. We train FII-CenterNet on the original 20 categories when compare with other methods. The performance is shown in (a). For the four concerned traffic object categories, FII-CenterNet achieves 86.9%(bicycle), 88.2%(car), 86.7%(motorbike) and 84.7%(person), respectively. If it is trained only on the four concerned traffic object categories, the mAP can be improved from 86.2% to 86.6%. The performance is shown in (b).

CenterNet achieves a **2.0%** performance improvement.

The third section in Table IV. contains other object detection methods exploiting segmentation information. Our method still shows a significant performance improvement compared with them. MR-CNN [22] is based on Faster R-CNN and DES [23] is on the basis of SSD.

As for the four concerned traffic object categories, our method has the best performance on motorbike detection compared with other methods. It also outperforms most of the methods using small input size on the other three categories. SSD300+VSSA(Vertical) [46] achieves the best accuracy on bicycle detection, which benefits from multi-resolution feature learning module and the vertical spatial sequence attention (VSSA) module. DSSD513 [9] has the best performance on car detection and achieves 88.7% AP. The highest accuracy on person detection is 83.7%, while our method achieves 82.7% AP.

Although FII-CenterNet can not achieve the best performance on every category, it achieves the **best speed-accuracy trade-off** among all the methods compared in Table IV. The test results on VOC 2007 highlight the effectiveness of our method.

For fairness, FII-CenterNet is trained on the original 20 categories when compared with other methods. The performance can be improved, if it is trained only on the four concerned traffic object categories. The mAP can be improved from 86.2% to 86.6% due to that. The corresponding precision-recall curves are shown in Figure 8 (b).

V. Conclusion

In conclusion, aiming to improve the performance of the anchor-free detectors for traffic object detection, this paper proposes FII-CenterNet, which introduces the foreground information to eliminate the interference of the complex background information in traffic scenes. The foreground region proposal network is based on semantic segmentation, which is supervised by the segmentation label generated from the bounding box label. Midground is introduced as the transition between foreground and background, which can provide rich edge information of the objects. The detection accuracy is efficiently improved for the introduction of both foreground location and scale information, which is verified by the experimental results on KITTI validation set. The results on KITTI benchmark and PASCAL VOC 2007 demonstrate that our FII-CenterNet achieves the state-of-the-art performance in both accuracy and efficiency.

For the future research, the segmentation branch can be added into the network to deal with the traffic object segmentation task, which can simply on the basis of the foreground region proposals. Besides, the detection results can be further used in other ITS applications, such as traffic congestion detection [49].

Acknowledgment

The authors would like to thank Prof. Qiulei Dong (Institute of Automation, Chinese Academy of Sciences) and Dr. Xuan Li (Institute of Automation, Chinese Academy of Sciences) for their invaluable discussions and support for this paper.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Adv. neural inf. proces. syst.*, 2012, pp. 1097–1105.

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Lect. Notes Comput. Sci.*, 2016, pp. 21–37.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[4] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

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

[6] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv:1904.07850*, 2019.

[7] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Lect. Notes Comput. Sci.*, 2018, pp. 765–781.

[8] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv:1804.02767*, 2018.

[9] C. Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," *arXiv: 1701.06659*, 2017.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[12] H. Law, Y. Teng, O. Russakovsky, and J. Deng, "CornerNet-Lite: Efficient keypoint based object detection," *arXiv preprint arXiv:1904.08900*, 2019.

[13] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6568–6577.

[14] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 850–859.

[15] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point set representation for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9656–9665.

[16] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "DenseBox: Unifying landmark localization with end to end object detection," *arXiv:1509.04874*, 2015.

[17] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5182–5191.

[18] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9626–9635.

[19] Q. Wang, J. Gao, and Y. Yuan, "Embedding structured contour and location prior in siamesed fully convolutional networks for road detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 230–241, 2018.

[20] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, 2020.

[21] A. Shrivastava and A. Gupta, "Contextual priming and feedback for faster r-cnn," in *Lect. Notes Comput. Sci.*, 2016, pp. 330–348.

[22] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware U model," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1134–1142.

[23] Z. Zhang, S. Qiao, C. Xie, W. Shen, B. Wang, and A. L. Yuille, "Single-shot object detection with enriched semantics," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5813–5821.

[24] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2403–2412.

[25] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More deformable, better results," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9300–9308.

[26] H. Zhang, D. Kang, H. He, and F.-Y. Wang, "APLNet: Attention-enhanced progressive learning network," *Neurocomputing*, vol. 371, pp. 166 – 176, 2020.

[27] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[28] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[29] X. Chen*, K. Kundu*, and Y. Zhu, "3D object proposals for accurate object class detection," in *Adv. neural inf. proces. syst.*, 2015, pp. 424–432.

[30] X. Yu, W. Choi, Y. Lin, and S. Savarese, "Subcategory-aware convolutional neural networks for object proposals and detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2017, pp. 924–933.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *arXiv: 1412.6980*, 2014.

[32] F. Yang, W. Choi, and Y. Lin, "Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2129–2137.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[34] Y. Chen, L. Tai, K. Sun, and M. Li, "MonoPair: Monocular 3D object detection using pairwise spatial relationships," *arXiv: 2003.00504*, 2020.

[35] J. Ren, X. Chen, J. Liu, W. Sun, J. Pang, Q. Yan, Y.-W. Tai, and L. Xu, "Accurate single stage detector using recurrent rolling convolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2017, pp. 752–760.

[36] B. Yang, J. Yan, Z. Lei, and S. Li, "CRAFT objects from images," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 6043–6051.

[37] S. Zhang, X. Zhao, L. Fang, F. Haiping, and S. Haitao, "Led: Localization-quality estimation embedded detector," in *Proc. Int. Conf. Image Process.*, 2018, pp. 584–588.

[38] L. Fang, X. Zhao, and S. Zhang, "Small-objectness sensitive detection based on shifted single shot detector," *Multimedia Tools Appl.*, vol. 78, no. 4, pp. 13 227–13 245, 2019.

[39] J. Wei, J. He, Y. Zhou, K. Chen, Z. Tang, and Z. Xiong, "Enhanced object detection with deep convolutional neural networks for advanced driving assistance," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1572–1583, 2020.

[40] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3D bounding box estimation using deep learning and geometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5632–5640.

[41] Y. Xiang, W. Choi, Y. Lin, and S. Savarese, "Subcategory-aware convolutional neural networks for object proposals and detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2017, pp. 924–933.

[42] X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3D object proposals for accurate object class detection," in *Adv. neural inf. proces. syst.*, 2015, pp. 424–432.

[43] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3D object detection for autonomous driving," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2147–2156.

[44] Z. Cai, Q. Fan, R. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 354–370.

[45] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen, "RON: Reverse connection with objectness prior networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5244–5252.

[46] Y. Yuan, Z. Xiong, and Q. Wang, "VSSA-NET: Vertical spatial sequence attention network for traffic sign detection," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3423–3434, 2019.

[47] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, "Inside-Outside Net: Detecting objects in context with skip pooling and recurrent neural networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2874–2883.

[48] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: object detection via region-based fully convolutional networks," in *Adv. neural inf. proces. syst.*, 2016, pp. 379–387.

[49] Q. Wang, J. Wan, and Y. Yuan, "Locality constraint distance metric learning for traffic congestion detection," *Pattern Recognit.*, vol. 75, pp. 272 – 281, 2018.

Siqi Fan received his B.E. degree from the Shanghai Jiao Tong University, Shanghai, China, in 2019. He is currently working toward the master's degree at the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences as well as University of Chinese Academy of Sciences.

His research interests include computer vision and intelligent vehicles.

Fenghua Zhu received the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2008. He is currently an Associate Professor with the State Key Laboratory for Management and Control of Complex Systems, China.

His research interests are artificial transportation systems and parallel transportation management systems.

Shichao Chen received the M.S. degree in Control Theory and Control Engineering from Beijing Forestry University, Beijing, China, in 2013. From 2013 to now, he is a research assistant with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China. He is currently pursuing a Ph.D. degree in Computer Technology and Application with Faculty of Information Technology from Macau University of Science and Technology, Macau, China.

His main research interests are in Edge Computing and Predictive Maintenance.

Hui Zhang received her B.S. degree from the Beijing Jiaotong University in 2015. She is currently a Ph.D. candidate at the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences as well as University of Chinese Academy of Sciences.

Her research interests include computer vision, pattern recognition, and intelligent transportation systems.

Bin Tian received the B.S. degree from Shandong University, Jinan, China, in 2009 and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2014. He is currently an Associate Professor of the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences.

His current research interests include computer vision, machine learning, and automated driving.

Yisheng Lv received the B.E. and M.E. degrees from the Harbin Institute of Technology, Harbin, China, in 2005 and 2007, respectively, and the Ph.D. degree from the Chinese Academy of Sciences, Beijing, China, in 2010. He is currently an Associate Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. He is also with Qingdao Academy of Intelligent Industries, Qingdao, China.

His current research interests include traffic data analysis, dynamic traffic modeling, and parallel traffic management and control systems.

Fei-Yue Wang (S'87-M'89-SM'94-F'03) received his Ph.D. in Computer and Systems Engineering from Rensselaer Polytechnic Institute, Troy, New York in 1990. He joined the University of Arizona in 1990 and became a Professor and Director of the Robotics and Automation Lab (RAL) and Program in Advanced Research for Complex Systems (PARCS). In 1999, he founded the Intelligent Control and Systems Engineering Center at the Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, under the support of the Outstanding Oversea Chinese Talents Program from the State Planning Council and "100 Talent Program" from CAS, and in 2002, was appointed as the Director of the Key Lab of Complex Systems and Intelligence Science, CAS. In 2011, he became the State Specially Appointed Expert and the Director of The State Key Laboratory for Management and Control of Complex Systems. Dr. Wang's current research focuses on methods and applications for parallel systems, social computing, and knowledge automation. He was the Founding Editor-in-Chief of the International Journal of Intelligent Control and Systems (1995-2000), Founding EiC of IEEE ITS Magazine (2006-2007), EiC of IEEE Intelligent Systems (2009-2012), and EiC of IEEE Transactions on ITS (2009-2016). Currently he is EiC of China's Journal of Command and Control. Since 1997, he has served as General or Program Chair of more than 20 IEEE, INFORMS, ACM, ASME conferences. He was the President of IEEE ITS Society (2005-2007), Chinese Association for Science and Technology (CAST, USA) in 2005, the American Zhu Kezhen Education Foundation (2007-2008), and the Vice President of the ACM China Council (2010-2011). Since 2008, he is the Vice President and Secretary General of Chinese Association of Automation. Dr. Wang is elected Fellow of IEEE, INCOSE, IFAC, ASME, and AAAS. In 2007, he received the 2nd Class National Prize in Natural Sciences of China and awarded the Outstanding Scientist by ACM for his work in intelligent control and social computing. He received IEEE ITS Outstanding Application and Research Awards in 2009 and 2011, and IEEE SMC Norbert Wiener Award in 2014.