L4Net: An anchor-free generic object detector with attention mechanism for autonomous driving

Yanan Wu | Songhe Feng | Xiankai Huang | Zizhang Wu

School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
Beijing Technology and Business University, Beijing, China
ZongMu Tech, Shanghai, China

Correspondence
Songhe Feng, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China.
Email: shfeng@bjtu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 61872032; Beijing Municipal Natural Science Foundation, Grant/Award Numbers: 4202058, 9192008

Abstract
Generic object detection is a crucial task for autonomous driving. To devise a safe and efficient object detector, the following aspects are required to be considered: high accuracy, real-time inference speed and small model size. Herein, a simple yet effective anchor-free object detector named L4Net is proposed, which incorporates a keypoint detection backbone and a co-attention scheme into a unified framework, and achieves lower computation cost with higher detection accuracy than prior art across a wide spectrum of resource constrains. Specifically, the backbone utilizes Multi-scale Receptive-fields Enhancement module (MRE) to capture context-wise information, where the features of object scale and shape invariance are simultaneously considered. The co-attention scheme integrates the strength of both Class-agnostic Attention (CA) and Semantic Attention (SA), and explores the valuable features from low-level to high-level to generate more accurate prediction boxes. Compared with previous feature fusion strategy, multi-scale features are selectively integrated by fully exploiting the different characteristics of low-level and high-level features, which leads to a small model size and faster inference speed. Extensive experiments on four well-known datasets demonstrate the effectiveness of our method. For instance, L4Net achieves 71.68% mAP on KITTI test set, with 13.7 M model size at the speed of 149 FPS on NVIDIA TX and 30.7 FPS on Qualcomm-based device, respectively, which is 4x smaller and 2x faster than baseline model.

1 | INTRODUCTION

In recent years, autonomous driving has become a hot topic in object detection tasks. However, limited by computational constraints on embedded devices [1,2], a real-time generic object detector with high accuracy on low-power devices becomes a critical and challenging task for autonomous driving [3]. Generally, existing real-time object detectors can be broadly divided into two categories [4]: anchor-based methods and anchor-free methods. Most successful anchor-based object detectors [5–14] usually enumerate large amounts of potential object locations and then categorize each proposal into a corresponding class label. However, such a strategy not only increases the computation cost when calculating the IoU, but also introduces many hyper-parameters, such as box size and aspect ratio [15]. Therefore, the detectors using anchor-box cannot be deployed in the autonomous driving due to the high time cost. Meanwhile, a large model size of this kind of model is not suitable for embedded systems [16–19].

To overcome the drawback of the above anchor-based methods, the anchor-free detection methods [20–25] are proposed, which bypass the requirement of anchor boxes and predictions generated in a point(s)-to-box style, and have drawn a lot of attention in computer vision community. Among these methods, one of the most representative method is CenterNet proposed by Zhou et al. [20], which models an object as a centre point of its bounding box and regresses to all other object properties (size and orientation) in each position. The centre point-based approach is end-to-end differentiable and does not require any post-processing, hence it is simpler and faster. However, such detection method in a per-pixel prediction fashion relies heavily on the accuracy of keypoints selection and the edge details of the object. For example, in Figure 1 (c), the object parts (the front wheels and the reflectors) that are hard to distinguish tend to be regrettably
ignored by CenterNet [20], which leads to the inaccuracy of prediction boxes localization, so as to further hinders efficient and safe driving.

To alleviate the above-mentioned issues, we propose L4Net, a single and unified network composed of a keypoint detection backbone and a co-attention scheme. For keypoint detection backbone, we basically follow the CenterNet [20] but make modifications to CenterNet by taking advantage of multi-scale receptive-fields to extract object scales and shape invariance since efficiently handling these complicated variations is crucial for improving the localization ability in object detectors based on keypoints estimation. In addition, the co-attention scheme incorporates class-agnostic attention and semantic attention to fusion low-level and high-level features jointly on top of the modified CenterNet, which is the key contribution of our method. Compared with previous works [5,6,8,26] that normally ignore the coefficient interaction of these two kinds of features and directly integrate multi-level features without distinction, for our model design, class-agnostic attention contributes more to focus on shallow features such as contours, texture and colors under the class-agnostic setting, while semantic attention lays emphasis on containing context-aware information to assign the object with correct labels [27]. Figure 1 (d) demonstrates the effectiveness of our proposed method by considering the fusion of shallow and abstract semantic features. Seeing that co-attention scheme can focus on more discriminative features meanwhile filter out redundant information, our method can apply fewer conv-layers and dense connection to learn valuable features, which makes our model achieve speed-accuracy trade-off with small model size. In short, the contributions are summarized as follows:

1. We propose L4Net, a completely anchor-free framework for object detection. L4Net jointly train the keypoint-based detection backbone and the co-attention scheme. In other words, The backbone utilizes multi-scale receptive-fields to capture object-level scale and shape invariance features. And the co-attention scheme integrates valuable features by class-agnostic attention and semantic attention to generate more accurate prediction boxes.

2. The proposed model achieves the speed-accuracy trade-off with small model size on the four well-known benchmark datasets (KITTI, PASCAL VOC 2007, MS COCO and BDD). Extensive experiments demonstrate the effectiveness of our method and practicability for autonomous driving applications.

2 | RELATED WORK

2.1 | Real-time object detectors

The real-time generic object detectors can be broadly divided into two categories: anchor-based and anchor-free methods. For anchor-based detectors [2,3,5–19], they are commonly classified into two-stage detectors and one-stage detectors. In two-stage detectors, the first stage generates large amounts of object proposals, and the second stage classifies the proposals as well as refines the coordinates with the anchor boxes. Such pipeline is first demonstrated its effectiveness by R-CNN [13] and is widely used in later two-stage detectors [16]. For instance, Qin et al. [16] present a lightweight backbone and adopt an extremely efficient design in the detection head and Region Proposal Network (RPN), which achieving superior detection accuracy. Compared with two-stage methods, the one-stage detectors skips object proposal generation and predicts bounding boxes and class scores in one evaluation, such as SSD [6], YOLO [4,7,8] and RefineDet [11]. When coupled with small backbone network, lightweight detector
Tiny-DSOD [12] introduces two innovative blocks: depthwise dense blocks (DDB) and depthwise feature pyramid networks (D-FPN). However, large amounts of hyper-parameters are required to define anchor boxes in these methods. Besides, exhaustive computation cost for IoU calculation can lead to large model size and slow inference speed.

Recently, to overcome the drawbacks mentioned above, the recent anchor-free object detection methods [20–25] have drawn a lot of attention where the prediction box is generated in a point(s)-to-box form. And the anchor-free object detectors can be roughly categorized into two groups: one is based on the anchor-point detection, which encodes and decodes object bounding boxes as anchor points with corresponding point-to-boundary distances, like FCOS [21], FoveaBox [22] and GA-RetinaNet [23]. In FCOS, all anchor points in a ground truth bounding box are used to predict the bounding boxes and the low-quality detected bounding boxes are suppressed by the proposed “centre-ness” branch; the other is based on keypoint detection, which predicts the locations of several keypoints (e.g., corners, centre or extreme points) of the bounding box, and groups those keypoints to form a box, such as CornerNet [24], ExtremeNet [25] and CenterNet [20]. The key step of CornerNet is to recognize which keypoints belong to the same instance and grouping them correctly. Herein, we mainly refer to CenterNet proposed by Zhou et al. [20], due to the fact that the principles of CenterNet are relatively easy to understand and very efficient since it models an instance as a single point without any grouping scheme to separate different instances.

### 2.2 | Attention Mechanism

Attention mechanism [27–35] has been successfully applied to various tasks such as machine translation, neural image/video captioning and object detection since it could assign greater weights to highly responsive feature maps. Bahdanau et al. [28] propose a network with attention to correlate the degree of association of each word in the output sequence with a particular word in the input sequence for machine translation. Xu et al. [29] propose the first visual attention method in image captioning and use both ‘hard’ and ‘soft’ attention mechanisms to focus on different parts of the image for each word in the output sequence. As for object detection, Hu et al. [30] propose a ‘Squeeze-and-Excitation’ block with adaptively weighting channel-wise feature in terms of corresponding responses. Perreault et al. [31] use the segmentation maps as a self-attention mechanism to weight the feature map used to produce the bounding boxes, decreasing the signal of non-relevant areas. Also, Zhou et al. [34] propose semantic self-attention module to effectively recognize foreground regions and suppress backgrounds.

Compared with existing methods [9,12,16,34] that directly utilize multi-level features without distinction, which leads to slower inference speed and sub-optimal results due to the distraction from redundant details, we propose an efficient co-attention scheme to obtain more accurate prediction boxes and faster inference speed by selectively integrate high-level and low-level features.

### 3 | PROPOSED METHOD

#### 3.1 | Overview

The overview architecture of our proposed L4Net is shown in Figure 2, which consists of keypoint detection backbone, co-attention scheme and detection head. For keypoint detection backbone design, we enhance the shallow features representation ability via steam block and utilize multi-scale receptive-fields to extract object-level shape-invariant features for accurately locating keypoint. The co-attention scheme contains two modules: class-agnostic attention and semantic attention, which are used to capture object boundary details and global context-aware information from low-level and high-level features, respectively. Through the fusion of these two attention, our model can filter out the distraction of background information and generate more accurate prediction boxes. For detection head, we basically follow the CenterNet [20].

#### 3.2 | Backbone design for keypoint detection

To improve the inference speed of object detector and match the input resolution to the backbone network, the L4Net uses small input resolution of 512×512 pixels. We observe that object spatial details tend to be ignored in existing methods [16,20], which leads to inaccuracy of prediction boxes localization. Therefore, we put more emphasis on absorbing more shallow features into our backbone network to boost recognizing object details such as edges and textures. In addition, we utilize multi-scale receptive-fields to capture more context-wise information and encode long-range relationship among pixels, which is vital for the localization subtask, especially for the localization of large objects.

Based on the above insights, we build a lightweight backbone based on residual block [10] to effectively detect keypoint. The network structure is shown in Table 1. Motivated by Pelee [3] and Inception-v4 [26], we first add a cost-efficient steam block after the input layer. The structure of stem block is shown on Figure 3 (Left). The steam block consists of a convolution layer with 7 × 7 convolution kernel at stride two to retain as much information as possible about the original image, and a concatenation module from three parallel branches to obtain complex feature maps at different scales and reduce computation cost: 3 × 3 at stride two maxpool, three consecutive convolution layers with 1 × 1 × 32, 3 × 3 × 32 at stride two and 1 × 1 × 64 convolution kernels, 3 × 3 × 64 at stride two convolution layer. Following the steam block, multi-residual blocks are stacked to enhance feature representations ability. The number of stacked residual blocks for different resolution is set in terms of the size distribution of object in the dataset. For
**FIGURE 2**  The overall architecture of L4Net. L4Net backbone is based on residual block and multi-scale receptive-fields enhancement module to obtain the feature representation of scale and shape invariance for more accurate locating keypoint. The channel self-attention module captures global semantic information for high-level features, the spatial self-attention module captures object boundary details for low-level features. The fusion of two attention modules not only improves the accuracy of the prediction box, but also reduces the computation overhead.

**TABLE 1** Architecture of the L4Net backbone networks

| Stage   | Layer | KSize | Stride | Repeat | Output Shape |
|---------|-------|-------|--------|--------|--------------|
| Image   | Input | -     | -      | -      | 512 × 512 × 3 |
| Stage0  | Steam | -     | 4      | 1      | 128 × 128 × 64 |
| Stage1  | Layer1| 3x3   | 3      | 3      | 128 × 128 × 64 |
|         | MRE   | -     | 1      | 1      | 128 × 128 × 128 |
| Stage2  | Layer2| 3x3   | 2      | 7      | 64 × 64 × 96  |
|         | MRE   | -     | 1      | 1      | 64 × 64 × 128 |
| Stage3  | Layer3| 3x3   | 2      | 3      | 32 × 32 × 128 |
|         | MRE   | -     | 1      | 1      | 32 × 32 × 128 |
| Detection layer | Centre point channel: C size channel: 2xC offset channel: 2 |
| FPS     | 149   |
| Weights/M | 3.6   |

example, there are a total of 34,856 object boxes in the KITTI dataset, among which 17.7% were small object (area < 32²), 53.7% were medium object (32² < area < 96²), and 28.6% were big object (96² < area). Therefore, we can increase the number of residual blocks at small feature maps for medium and large object predictions.

In addition, local invariance is very significant for detection task. However, the existing models generally directly stack the convolution and pooling layers in a bottom-to-up style, which may not be able to deal with some complex scenes effectively. Here, inspired by Scale-invariant feature transform (SIFT) [36], we devise a multi-scale receptive-fields enhancement module to obtain the feature representation of scale, position and shape invariance. The SIFT is a vision algorithm to search extreme points on spatial scale and gain their location, rotation and scale invariants, where it is usually used to extract local features of images.

The SIFT proposed the difference-of-Gaussian (DOG) scale-space and image Gaussian pyramid [36]. The DOG

**FIGURE 3** Left: Structure of stem block. Right: Multi-scale Receptive-fields Enhancement Module
scale-space are processed by several different Gaussian kernel functions at the same resolution; and the Gaussian pyramid are generated by the Gaussian function to blur and downsamples the image with different resolutions. Similar to DOG, we adopt dilated convolution with various dilation rates (3, 5, 7) to acquire feature maps of same scale but different receptive-fields. Afterwards, the feature maps of different dilated convolutional and a $1 \times 1$ dimension reduction feature by cross-channel are concatenated for each layer. Finally, we change output channels of the MRE by the $1 \times 1$ convolution to fit input channels of the next layer. Similar to image Gaussian pyramid, we take Layer 1, Layer 2 and Layer 3 of L4Net backbone to obtain multi-scale feature maps. In a word, multi-scale receptive-fields enhancement module is shown in Figure 3 (Right).

### 3.3 Attention mechanism

In this subsection, we elaborate the co-attention scheme to filter out background distraction information significantly and focus more on valuable features. The co-attention scheme is composed of semantic attention and class-agnostic attention, which capture context-aware information and object boundaries details from high-level and low-level features, respectively.

#### 3.3.1 Semantic attention

We devise semantic attention after multi-scale receptive-fields (MREs) to assign greater weights to channels which shows more active response to semantic interest of high-level feature. Since each convolutional layer performs as a pattern detector, each channel of the feature map generated by convolutional layer is a response activation of the corresponding semantic. Applying semantic attention in a channel manner can be regarded as a process of selecting the object with correct labels.

For semantic attention (SA), we define high-level features $f_\delta^b \in \mathbb{R}^{W \times H \times C}$ as $f_\delta^b = \{f_\delta^1, f_\delta^2, \ldots, f_\delta^C\}$, where $f_\delta^b \in \mathbb{R}^{W \times H}$ is the $i$-th slice of $f_\delta^b$, where $W$ and $H$ represent width and height of feature map and $C$ represents the channel number. First, to exploit correlations between channels, we apply global average pooling to obtain channel statistics vector $\mathbf{v}^b \in \mathbb{R}^C$, where the channel vector $\mathbf{v}^b$ is generated by shrinking $f_\delta^b \in \mathbb{R}^{W \times H}$ in terms of its spatial dimensions $W \times H$ and $\mathbf{v}^b$ is calculated by

$$v_i^b = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} f_{\delta}^b(i, j).$$

Then we use two fully connected layers to extract correlations between channels, as shown in Figure 4. Motivated by Hu et al. [30], to aid generalization and reduce computation cost, we encode the channel vector with reducing the feature dimension to $1/16$ of the input in the first FC layer, then activating it with ReLU and recovering to the original dimension through the second fully connected layer. Finally, the normalized channel vector weights mapped to $[0, 1]$ are obtained by a sigmoid operation following

$$SA = \sigma_1(f_{c2}(\delta(f_{c1}(\mathbf{v}^b, W_1)), W_2)),\ (2)$$

where $W_1$ and $W_2$ refer to parameters in semantic attention, $\sigma_1$ refers to sigmoid operation, $f_{c1}$ and $f_{c2}$ represent fully connected layers, $\delta$ is the activation function ReLU. The final output $\tilde{f}^b \in \mathbb{R}^{W \times H \times C}$ is calculated by weighting the high-level features with semantic attention.

$$\tilde{f}^b = SA \odot f^b;\ (3)$$

where $\odot$ denotes the outer product of weighted channel vectors and high-level features.

#### 3.3.2 Class-agnostic attention

Considering all spatial positions equally resulting in sub-optimal results with distraction of background details, we design class-agnostic attention to focus more on the object boundaries details for generating more accurate prediction boxes.

For class-agnostic (CA), We define low-level feature maps as $f_i \in \mathbb{R}^{W \times H \times C}$. The $\mathbb{R} = \{(x, y) \mid x = 1, \ldots, W; y = 1, \ldots, H\}$ represents the set of spatial positions, where $p = (x, y)$ denotes the spatial coordinate of low-level features. As shown in Figure 5, class-agnostic attention is generated through the following steps: At first, for reducing computation cost and obtaining rich global information, we, respectively, approximate the $k \times k$ kernel to ‘1 $\times$ k’ + ‘k $\times$ 1’ and ‘k $\times$ 1’ + ‘1 $\times$ k’ to capture different spatial concerns as $C_1 \in \mathbb{R}^{W \times H \times 1}$ and $C_2 \in \mathbb{R}^{W \times H \times 1}$ for low-level features. Then, the class-agnostic attention feature $CA \in \mathbb{R}^{W \times H \times 1}$ can be generated by sigmoid operation normalizing the addition of $C_1$ and $C_2$ following
and optimizing the overall goal by Adam from scratch. Finally, the experiment is conducted on NVIDIA TITAN X with CUDA 8.0 and deployed on Qualcomm-based low-power embedded devices by Snapdragon Neural Processing Engine (SNPE) [37]. The detailed experimental results are introduced in the following subsection.

4.2 The results on KITTI benchmark

First, our model is conducted from scratch on KITTI dataset, which contains 7481 training images collected from real-world autonomous driving scenarios [38]. Meanwhile, Half of the 7481 images are randomly selected as the training set, and the rest as the validation set. The detection metrics are tested on the validation set. During the training, We set a mini batch size of 64. The initial learning rate is set to $5 \times 10^{-4}$, and divided by a factor of 10 at 90 and 120 epochs, respectively. Our training stops at 200 epochs. Moreover, since our training dataset is small, we use heavy data enhancements including random flip, random scaling, cropping and colour jitter to prevent model overfitting and improve robustness. However, in the inference phase, data enhancements will increase too much inference time and so not be used.

The detection results of the proposed method and other methods on validation set are reported in Table 2. The detection speed of the proposed model, L4Net, is 49 fps faster than CenterNet [20], which has the fastest inference speed among the previous object detectors, despite the mAP of L4Net outperforming that of CenterNet [20] by more than 6.5%. Furthermore, the KITTI dataset consists of a strongly wide set of 1242 × 375 images. To avoid the disappearance of small objects, the input size of our model is resize to 1200 × 300. It is noticed that the mAP of L4Net with a large input size of 1200 × 300 is slightly better than L4Net (original), but the inference speed of L4Net-large is 53 fps slower. We observe that the large input image can indeed detect more small object. However, in practice, the autonomous driving task generally needs to accurately detect the objects within 10 m around the vehicle. Therefore, it is unnecessary to use large input images and extra computational overhead to detect small, meaningless objects at a distance.

In addition, although the mAP of L4Net is 16.34% lower than that of FCOS [21], which has the highest accuracy among the previous anchor-free detection models, it is noteworthy that the inference speed of the proposed method is 9x better than FCOS [21]. Our method achieves trade-off between accurate and speed while FCOS spends too much computational cost (i.e. large input size and multi-scale prediction) to improve accuracy, which is unsafe due to high latency for autonomous driving application.

Table 3 shows the inference speed and model size of our L4Net and other methods on real device Qualcomm 820. The speed is calculated by the average time that the benchmark tool processes 50 images. According to Table 3, L4Net runs approximate 2x faster with 4x smaller model size than CenterNet [20] on real device, while L4Net and CenterNet

$$C_1 = \text{conv}_2 \left( \left( \text{conv}_1 \left( f^l, W_1^l \right), W_2^l \right) \right),$$

$$C_2 = \text{conv}_1 \left( \left( \text{conv}_2 \left( f^l, W_1^l \right), W_2^l \right) \right),$$

$$CA = \sigma_2 \left( C_1 + C_2 \right),$$

where $W_1$ and $W_2$ represent parameters in class-agnostic attention, $\sigma_2$ is sigmoid function, $\text{conv}_1$ and $\text{conv}_2$ refer to $1 \times k$ and $k \times 1$ convolution layer, respectively. Meanwhile, $k$ is hyper-parameter, which is set 9 on our model. The final output $\sim \sim f^l \in \mathbb{R}^{W \times H \times C}$ of the block is obtained by weighting $f^l$ with CA.

$$\sim \sim f^l = CA \odot f^l,$$

where $\odot$ represents the outer product of weighted class-agnostic attention feature and low-level features.

4 EXPERIMENTS

4.1 Implementation details

To evaluate the performance of our proposed method, we use 15 state-of-the-art object detectors for comparative studies and four widely used real-time object detection metrics (mean average precision [mAP], average recall [AR], speed and model size) are also utilized to evaluate each comparing method. Here, all detection metrics are the average of 10 experiments. In addition, we conduct comprehensive ablation study and compare our proposed L4Net with its degenerated model to demonstrate the effectiveness of each component in our model.

During the process of experiments, most training strategies follow CenterNet, such as data augmentation, input size, classification and regression of prediction layers, loss function (focal loss [17] for classification and L1 loss for localization)
have similar speed on TITAN X GPU. As shown in the second column of Table 2, mode size of FCOS [21] is 2x and 9x larger than CenterNet and L4Net, respectively, which is unfeasible embedded system deployment, although FCOS has higher accuracy (see Table 2). In addition, the model size of Tiny-DSOD [12] is 4.2 M, which is 3x smaller than L4Net, but it turns out that Tiny-DSOD and L4Net have similar inference speed. This is because L4Net only generates a prediction box for each object without complicated NonMaximum Suppression (NMS) [5].

4.3 | The results on PASCAL VOC2007 benchmark

Next our detector is evaluated on the combination of VOC2007 and VOC2012 trainval datasets, and tested on the VOC2007 test set [39]. A mini batch size of 128 is set. The initial learning rate is set to 5 × 10⁻⁵ in training process, since we observed that loss is oscillating and not smooth with a large learning rate from scratch. And we divide by a factor of 5 at 60 and 100 epochs, respectively. The iteration stops at 160 epochs as the loss tends to decline slowly in training. Other settings are the same as on KITTI dataset.

Table 4 summarizes the detection metrics on VOC2007 test set. The results of the upper part come from state-of-the-art anchor-based detectors, and the results of the lower part come from anchor-free detectors. In terms of mAP, Our L4Net achieves significant accuracy 73.1%, which outperforms most anchor-based detection models, except RetinaNet. However, our L4Net is 18x faster in fps than RetinaNet. When comparing our L4Net with the state-of-the-art anchor-free detectors, the accuracy drops. But the L4Net demands much less computing cost and has faster inference speed. For instance, FoveaBox [22] is 8.4% higher in accuracy than L4Net, while with more than 16x slower inference speed. The above comparisons reveal that L4Net has much better trade-off between inference speed and detection accuracy, which is of great use for resource-restricted devices.

4.4 | The results on MS COCO benchmark

We also evaluate the performance of our proposed model on the MS COCO test-dev set, which is composed of 118k training images, 5k validation images and 20k images [20,40].
The batch size of training is set to 32, and the model is trained with $5 \times 10^{-4}$ learning rate for the first 90 epochs, then $5 \times 10^{-5}$ and $5 \times 10^{-6}$ for another 30 and 60 epochs, respectively.

Table 5 reports the detection results on the COCO test-dev set. L4Net achieves 31.2% mAP, which is more than the anchor-based RefineDet [11] and Pelee [2], and even outperforms the lightweight detector Tiny-DSOD [12].

Compared with EfficientDet-D0 [19] and SNIPER [14], our L4Net achieves similar accuracy with up to 2x faster inference speed and 4x fewer input size. Besides, L4Net has significant improvement compared with all the anchor-free methods listed in Table 5. For example, L4Net has 19x faster speed than CornerNet [24]. These comparisons verify that L4Net is the first real-time yet accurate anchor-free detector for resource-restricted object detection usages. Meanwhile, we show some qualitative results on the KITTI, VOC 2007 and the COCO test-dev in Figure 6 respectively. Our method works well with the occlusions, interclass interference and cluttered background.

**4.6 | Ablation study on KITTI**

Keypoint detection backbone improves localization ability of object centre point. All variants results in ablation studies are based on a compressed CenterNet with 63.21% mAP (Table 7(a)) as our baseline which reduces channels of convolution kernel than official CenterNet to generate small model size. We first apply the steam block after input layer of CenterNet. Result is reported in Table 7 (b). It is evident that steam block can offer a significant improvement (up to 0.96% mAP) over the baseline without adding too much parameters. As for Multi-scale receptive-fields enhancement (MRE) module, the accuracy reached 66.43% mAP (Table 7 (c)). Furthermore, we do a statistic on overlapping ratio (distance threshold $\leq \sqrt{2}$) of object centre point between ground truth and the proposed method is 42.6%, which improves 3.7% than CenterNet. Therefore, it turns out that the accuracy of object centre point localization can be improved by capturing features of object scales and shape invariance according to MRE.

| Model          | Backbone | Input size | FPS  | mAP  |
|----------------|----------|------------|------|------|
| SSD300* [6]   | VGG-16   | 300 × 300  | 46   | 75.9 |
| YOLOv2 [7]    | DarkNet-19 | 416×416    | 67   | 70.5 |
| RefinedDet [11]| VGG-16   | 320 × 320  | 40.3 | 80.0 |
| EfficientDet-D0 [11] | EfficientNet | 512 × 512  | 62.5 | 75.0 |
| Tiny-DSOD [12] | G/32-48-64-80 | 300 × 300  | 105  | 72.1 |
| RefinedDet-V3-Tiny [8] | Dark-Tiny | 416×416    | 67   | 54.0 |
| RetinaNet [17] | ResNeXt-101 | 800 × 800  | 8.9  | 80.8 |
| MobileNet-SSD [6, 8] | MobileNetv2 | 300 × 300  | 73   | 68.0 |
| Pelee [2]     | PeleeNet  | 304 × 304  | 75   | 70.9 |
| **Anchor-Free Detectors** |          |            |      |      |
| ExtremNet [25] | Hourglass-104 | 511 × 511  | 2.8  | 80.1 |
| CornerNet [24] | Hourglass-104 | 511 × 511  | 2.9  | 80.7 |
| FoveaBox [22]  | ResNeXt-101 | 800 × 800  | 10.2 | 81.5 |
| FCOS [21]      | ResNeXt-101 | 1024 × 800 | 9.7  | 81.9 |
| CenterNet [20] | ResNet-18   | 384 × 384  | 142  | 70.2 |
| L4Net          | L4Net      | 300 × 300  | 168  | 73.1 |
| L4Net(acc)     | ResNeXt-101| 512 × 512  | 24   | 82.3 |

**4.5 | The results on BDD benchmark**

Finally, we evaluate the performance of the proposed model and other models on the BDD test set, which is the latest published autonomous driving dataset with the largest scale and most diverse content. And the BDD dataset consists of ten classes: bike, bus, car, motor, person, rider, traffic light, traffic sign, train and truck. The ratio of training, validation and test set is 7:1:2. In L4Net training, the batch size is 64 and the initialization learning is $5 \times 10^{-4}$.

Table 6 shows the experimental results on BDD test set. L4Net is 149 fps faster than the CenterNet [20], which has the highest operation speed among the previous studies, despite the accuracy of L4Net outperforming that of CenterNet [20] by 0.3 mAP. In addition, compared with FoveaBox [22], which has the highest accuracy among the previous real-time detectors, the performance of L4Net with a 1024 × 800 input resolution and ResNeXt backbone in the last row of Table 6 shows a better mAP of 2.2 and faster operation speed of 8.0 fps, and consequently, L4Net outperforms FoveaBox [22] in terms of the accuracy and detection speed.

**4.6 | Ablation study on KITTI**

Keypoint detection backbone improves localization ability of object centre point. All variants results in ablation studies are based on a compressed CenterNet with 63.21% mAP (Table 7(a)) as our baseline which reduces channels of convolution kernel than official CenterNet to generate small model size. We first apply the steam block after input layer of CenterNet. Result is reported in Table 7 (b). It is evident that steam block can offer a significant improvement (up to 0.96% mAP) over the baseline without adding too much parameters. As for Multi-scale receptive-fields enhancement (MRE) module, the accuracy reached 66.43% mAP (Table 7 (c)). Furthermore, we do a statistic on overlapping ratio (distance threshold $\leq \sqrt{2}$) of object centre point between ground truth and the proposed method is 42.6%, which improves 3.7% than CenterNet. Therefore, it turns out that the accuracy of object centre point localization can be improved by capturing features of object scales and shape invariance according to MRE.
Co-attention scheme regresses more accurate prediction boxes. Next, we further apply the co-attention scheme on the top of the keypoint detection backbone, so that the proposed model can focus on different-levels valuable features. Seeing that co-attention scheme consist of semantic attention (SA) and class-agnostic attention (CA), Table 7 (d) indicates that semantic attention pushes our variant model to a 68.67% mAP, while Table 7 (c) means that class-agnostic attention without SA improves 1.12% mAP with merely 0.2 M parameters augment compared with Table 7 (c). It is clear that
either of SA or CA effectively improves detection accuracy by focusing on the most representative feature.

At last, The semantic attention and class-agnostic attention are composed of the complete L4Net (Table 7 (f)), where the co-attention improves mAP by 5.25% with negligible increase on parameters than Table 7 (c). With our co-attention scheme, the inaccurate issue of prediction boxes at the object edge details is effectively alleviated and its effectiveness also further is confirmed by the error analysis protocol provided in the study Lin et al. [40] in Figure 7. Each sub-image depicts a series of precise-recall curves with different settings, where the area under each curve is displayed (brackets) in the images. A larger IoU reveals that the coordinates of the prediction boxes are closer to ground truth. The overall AP and the medium object AP at IoU = 0.75 of L4Net improve by 4.4% and 11.4% compared with CenterNet, which indicates our model performs better in prediction boxes localization ability. We also visualize the detection results in Figure 6. These figures clearly show that this is consistent with our initial motivation to design co-attention scheme for different level features.

5 CONCLUSION AND FUTURE WORK

We have proposed an anchor-free generic object detector named L4Net that achieved the best speed-accuracy trade-off with small model size for autonomous driving. For keypoint detection backbone, we utilized multi-scale receptive-fields to capture object scales and shape invariance. The co-attention scheme consists of class-agnostic attention and semantic attention which captured object spatial details and context-aware information from low-level and high-level features, respectively. Compared with the baseline, the proposed L4Net model, respectively, improves the mAP by 6.51, 2.9, 1.4, 0.3 for the KITTI, PASCAL VOC, MS COCO and BDD datasets, respectively, with 13.7 M model size at the speed of 149 FPS on NVIDIA TX and 30.7 FPS on Qualcomm-based device. As a result, the proposed method significantly improved the performance of object detection system for autonomous driving applications.

Furthermore, although our model achieves the satisfactory trade-off between accuracy and speed for autonomous driving, some details are easily ignored when we maintain the fast speed. And our model is poor performance when it is suffered from adverse weather (rainy, foggy), light (in and out of tunnels), and large obstructions. Therefore, in the future work L4Net will be improved by introducing the effective network structure design, such as geometric graph-based multi-view learning.

ACKNOWLEDGEMENT

This study was funded in part by National Natural Science Foundation of China (No. 61872032, No. 62072027), Beijing Municipal Natural Science Foundation (No. 4202058, No. 4202057), the National Key Research and Development Project under Grant 2018AAA0100300, and in part by the Fundamental Research Funds for the Central Universities (2020YJS036).

REFERENCES

1. Jiao, L., Zhang, F., Liu, F., et al: A survey of deep learning-based object detection. IEEE Access. 7, 128837–128868 (2019)
2. Wang, R.J., Li, X., Ling, C.X.: Peele: a real-time object detection system on mobile devices. Adv. Neural Inf. Process. Syst., 1963–1972 (2018)
3. Choi, J., Chun, D., Kim, H.: YOLOv3: ‘Gaussian: An accurate and fast object detector using localization uncertainty for autonomous driving’ (2019). arXiv: Computer Vision and Pattern Recognition
4. Wu, X., Sahoo, D., Hoi, S.C.H.: Recent advances in deep learning for object detection. Neurocomputing (2020)
5. Lin, T.Y., Dollár, P., Girshick, R.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
6. Liu, W., Anguelov, D., Erhan, D.: Ssd: single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37 (2016)
7. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
8. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement’ (2018). arXiv preprint arXiv:1804.02767
9. Sandler, M., Howard, A., Zhu, M.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
10. Xu, K., Zhang, X., Ren, S.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Zhang, S., Wen, L., Bian, X.: Single-shot refinement neural network for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4203–4212 (2018)
12. Li, Y., Li, J., Lin, W: ‘Tiny-iod: lightweight object detection for resource-restricted usages (2018). arXiv preprint arXiv:1807.11013
13. Girshick, R., Donahue, J., Darrell, T.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
14. Singh, B., Najibi, M., Davis, L. S.: SNIPER: efficient Multi-Scale Training’. neural information processing systems, pp. 9310–9320 (2018)
15. OksuZ, K., Cam, B.C., Kalkan, S., et al: Imbalance problems in object detection: a review. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
16. Qin, Z., Li, Z., Zhang, Z., et al: ThunderNet: towards real-time generic object detection on mobile devices. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 6718–6727 (2019)
17. Lin, T.Y., Goyal, P., Girshick, R., et al: Focal loss for dense object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
18. Zhou, X., Zhuo, J., Krahenbuhl, P.: Bottom-up object detection by grouping extreme and center points. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 850–859 (2019)
26. Szegedy, C., Ioffe, S., Vanhoucke, V., et al.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: 31st AAAI Conference on Artificial Intelligence (2017)

27. Lee, J.B., Rossi, R.A., Kim, S., et al.: Attention models in graphs: a survey'. ACM Trans. Knowl. Discov. Data. 13(6), 1–25 (2019)

28. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate'. (2014). arXiv preprint arXiv:1409.0473

29. Xu, K., Ba, J., Kiros, R.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)

30. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

31. Perreault, H., Bilodeau, G. A., Saunier, N., et al.: SpotNet: self-attention multi-task network for object detection (2020). arXiv preprint arXiv:2002.05540

32. Woo, S., Park, J., Lee, J.Y., et al.: Cham: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision, pp. 3–19 (2018)

33. Zhang, X., Wang, T., Qi, J., et al.: Progressive attention guided recurrent network for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 714–722 (2018)

34. Chen, L., Zhang, H., Xiao, J.: 'SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning'. Comput. Vis. Pattern Recogn., 6298–6306 (2017)

35. Zhou, C., Wu, M., Lam, S.: 'SSA-CNN: Semantic self-Attention CNN for pedestrian detection (2019). arXiv preprint arXiv:1902.09080

36. Lowe, G.: SIFT-the scale invariant feature transform. Int. J. Comput. Vis. 2, 91–110 (2004)

37. Overview - Snapdragon Neural Processing Engine SDK’, https://developer.qualcomm.com/docs/snpe/overview.html, Accessed 5 April 2020

38. Andreas, G., Philip, L., Raquel, U.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2012)

39. Everingham, M., Van, G.L., Williams, C.K.I, et al.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. 88(2), 303–338 (2010)

40. Lin, T. Y., Maire, M., Belongie, S.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision, pp. 740–755 (2014)

How to cite this article: Wu Y, Feng S, Huang X, Wu Z. L4Net: An anchor-free generic object detector with attention mechanism for autonomous driving. IET Comput. Vis. 2021;15:36–46. https://doi.org/10.1049/cvi2.12015