A Deep Model for Joint Object Detection and Semantic Segmentation in Traffic Scenes

Jizhi Peng*, Zhixiong Nan*, Linhai Xu, Jingmin Xin and Nanning Zheng
Institute of Artificial Intelligence and Robotics
Xi'an Jiaotong University
Xi'an, Shaanxi, 710049 P.R. China
pengjz@stu.xjtu.edu.cn, {nzx2018, xlh, jxin, nnzheng}@xjtu.edu.cn

***Abstract*—Object detection and semantic segmentation are two fundamental techniques of various applications in the fields of Intelligent Vehicles (IV) and Advanced Driving Assistance System (ADAS). Early studies separately handle these two problems. In this paper, inspired by some recent works, we propose a deep neural network model for joint object detection and semantic segmentation. Given an image, an encoder-decoder convolution network extracts a set of feature maps, these feature maps are shared by the detection branch and the segmentation branch to jointly carry out the object detection and semantic segmentation. In the detection branch, we design a PriorBox initialization mechanism to propose more object candidates. In the segmentation branch, we use the multi-scale atrous convolution to explore the global and local semantic information in traffic scenes. Benefiting from the PriorBox Initialization Mechanism (PBIM) and Multi-Scale Atrous Convolution (MSAC), our model presents the competitive performance. In the experiments, we widely compare with several recently-proposed methods on the public Cityscapes dataset, achieving the highest accuracy. In addition, to verify the robustness and generalization of our model, the extension experiments are also conducted on the well-known VOC2012 dataset.**

***Index Terms*—traffic scenes, object detection, semantic segmentation**

I. Introduction

Object detection and semantic segmentation are two important tasks in the computer vision, serving as the fundamental technique support for many applications of autonomous driving car and advanced driving assistance system [1], [2]. In early years, the object detection and semantic segmentation are studied as two separate problems. The researches on joint object detection and semantic segmentation sprout from the works in [3]–[5]. Actually, object detection and semantic segmentation are two highly correlated tasks so that they could be mutually beneficial. Semantic segmentation could provide both global and local semantic information to the object detection (e.g., vehicles usually run on the roads rather on the sky), and object detection provides the prior knowledge to refine the semantic segmentation (e.g., the regions beneath a vehicle bounding box tend to be the roads). Considering the computational requirements, joint object detection and semantic segmentation often share the same feature extraction network to save the computational requirements. Therefore, it is meaningful and necessary to stride towards deeper study of joint object detection and semantic segmentation.

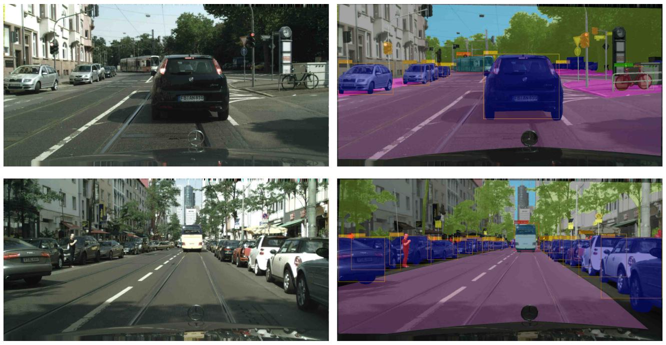

Fig. 1. Samples of joint object detection and semantic segmentation.

As shown in Fig. 1, the goal of joint object detection and semantic segmentation is to simultaneously detect objects in images and segment the image into semantic regions. To realize the joint object detection and semantic segmentation, many good models have been proposed [6]–[8]. BlitzNet [6] is an encoder-decoder network for joint object detection and semantic segmentation. It uses each decoder layer for multi-scale object detection and concatenates decoder layers to perform semantic segmentation. TripleNet [8] also takes the encoder-decoder network as the backbone network and uses multiple decoder layers for the object detection and semantic segmentation. These methods have achieved the impressive performance. One main advantage of these methods is that the skip-layer mechanism is involved in the encoder-decoder network, contributing to extracting the feature maps conveying both high-level and low-level information. Another advantage is that multiple-scale feature maps are simultaneously utilized for the detection and segmentation, which has been proved to exhibit higher performance than the model only using the single-scale feature map.

In this paper, besides utilizing the above mentioned advantages, we also consider two other important factors for joint object detection and semantic segmentation: 1) The object candidate proposal method is significant for object detection.

Jizhi Peng* and Zhixiong Nan* are co-first authors.

This work is supported by the National Natural Science Foundation of China (NO.61773312,61790563,61790562) and "the Fundamental Research Funds for the Central Universities".

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Previous methods usually generate object candidates by setting PriorBox on the multiple-scale feature maps using the predefined rule. Considering that the traffic scenes are highly dynamic and complex, we define a new PriorBox initialization mechanism to generate denser object candidates; 2) Semantic segmentation signals an overall understanding for an image, thus it is crucial to involve the global and local semantic information in the model. Previous methods often use the feature maps in the shallow layers of the network to extract the local semantic information and the feature maps in the deep layers to extract the global semantic information. However, the convolution kernel is often with small sizes (e.g., 3×3), which limits to explore the overall semantics since small convolution kernels obtain small receptive fields. Therefore, inspired by some works (e.g., [9]–[12]), we adopt the atrous convolution to enlarge the receptive fields.

Considering these two factors, we propose an encoder-decoder neural network model with the two branches that are respectively targeted for object detection and semantic segmentation. In the encoder-decoder backbone network, we adopt the skip-connection mechanism to fuse the feature maps of encoder layers and decoder layers. We also involve the Squeeze-and-Excitation [13] module in the skip-connection mechanism to improve the representation power of the feature maps. In the detection branch, we apply a self-defined PriorBox initialization mechanism for object candidate generation, and the object candidates are further processed by the classifiers to realize the detection. In the segmentation branch, we employ the multiple-scale atrous convolution on the deepest decoder layer to generate a feature map, which is concatenated with upsampled features of other decoder layers to form the final feature map for the semantic segmentation.

In the experiments, our proposed model is compared with several recent methods on the well-known public Cityscapes [14] dataset, achieving the best performance. In addition, to verify the generalization and robustness of our model, we also test our model on the VOC2012 [15] dataset which is not collected in the traffic scenes, and the results validate the effectiveness of our method.

The paper is organized as follows: Section 2 discusses related work, Section 3 presents our proposed method for joint object detection and semantic segmentation, Section 4 describes experiments, and Section 5 summarizes the paper.

II. Related works

A. Object detection

Object detection aims to classify and locate objects in an image, and the object detection methods are mainly divided into the two-stage method and the single-stage method.

Two-stage detection methods first generate a set of region proposals and then refine them by the classifier. R-CNN [16] is a classic object detection model. In order to reduce the excessive computational consumption in R-CNN, SPPNet [17] and Fast R-CNN [18] extract the features of the entire image, and then generate regional features through the spatial pyramid pool and RoIPooling layer, respectively. Faster R-CNN [19] proposes an end-to-end object detection architecture for the first time. It proposes a region proposal network(RPN) to improve the efficiency of the detector. Based on Faster R-CNN, Cascade R-CNN [20] proposes a multi-level detector through powerful cascade architecture. Compared with Faster R-CNN, Mask R-CNN [21] predicts instance segmentation by adding a mask branch, and the RoIPooling layer in Faster R-CNN is improved to RoIAlign layer to solve the problem of mis-alignment.

One-stage methods do not have the region proposal module. Instead, they directly classify and refine the pre-defined anchors. YOLO [22] and SSD [23] are the two earliest one-stage detection methods. Based on SSD, DSSD [24] uses the deconvolution to add extra context information to improve the detection accuracy. In order to solve the problem of the class imbalance during the training, RetinaNet [25] proposes the focal loss to down-weight the contribution of easy samples. In addition, it uses the feature pyramid network FPN [26] to form an encoder-decoder structure to enhance context information connection.

B. Semantic segmentation

The main goal of semantic segmentation is to predict which category each pixel in an image belongs to. FCN [27] is the first approach to adopt fully convolution network for semantic segmentation. Because multiple down-sampling layers are used in FCN's architecture, they usually have lower resolution in high-level feature maps. In order to solve the problem of low resolution caused by multiple downsamplings in FCN, DeepLabv1 [9] rebuilds the network architecture and uses the multiple layers of atrous convolution layers to expand the receptive field size. Besides, some works [28], [29] use the feature maps from earlier feature layers to compensate for the lower resolution of high-level features. SegNet [28] and UNet [29] are representative networks using the skip connection, which are called encoder-decoder architecture networks. Because objects have different scales, which require different context information, PSPNet [30] and Deeplabv3+ [12] are proposed to concatenate features of multiple receptive field sizes together for final prediction. PSPNet uses 4 parallel spatial pyramid poolings to receive information at multiple scales. Deeplabv3+ [12] concatenates the features from multiple atrous convolution layers with different dilation rates arranged in parallel.

C. Joint object detection and semantic segmentation

The goal of joint object detection and semantic segmentation is to carry out the object detection and semantic segmentation simultaneously. The idea of joint object detection and semantic segmentation originates from [3]–[5], which suggests that learning two tasks simultaneously might be better than learning each task alone. UberNet [31] integrates multiple visual tasks such as semantic segmentation and object detection into a single deep neural network. However, this architecture is not end-to-end. BlitzNet [6] is a real-time network for joint detection and semantic segmentation. It is

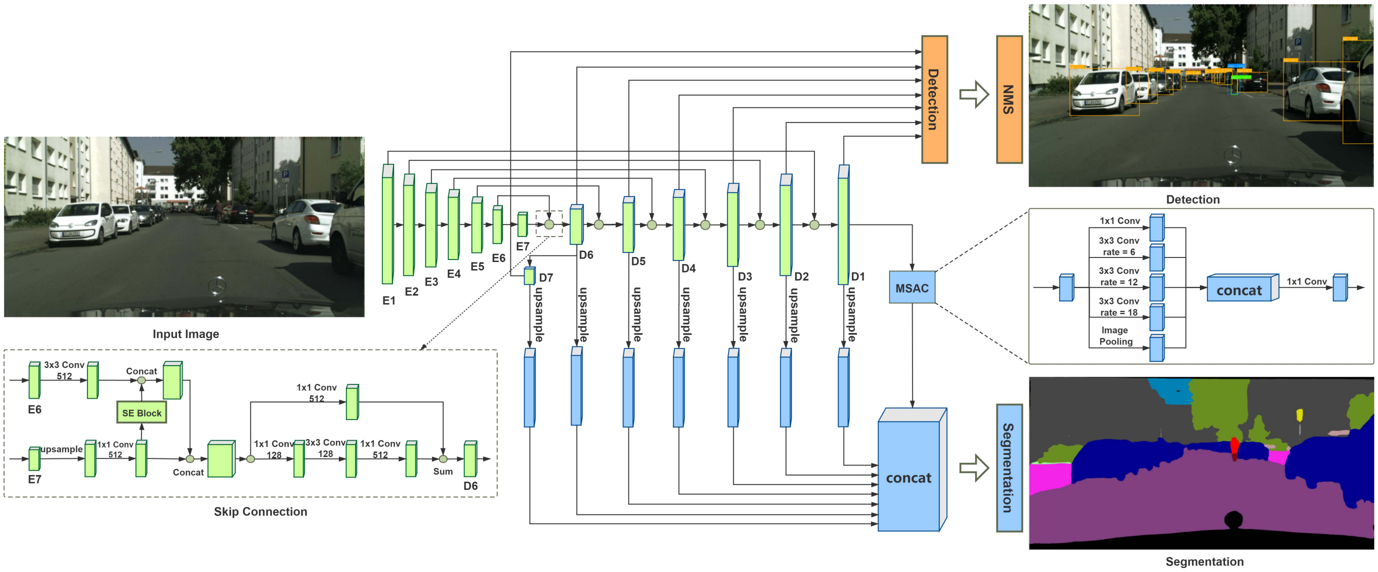

Fig. 2. The overview of our proposed method for joint object detection and semantic segmentation. The green color represents the encoder-decoder module, the orange color represents the object detection module, and the blue color represents the semantic segmentation module. "Skip connection" is the network to fuse the feature maps of the encoder layers and decoder layers. "MSAC" is the Multi-Scale Atrous Convolution network. For the convenience, the "upsample" in the figure means that the corresponding feature map is upsampled to the same scale. Layers of the decoder are simultaneously used for the detection and segmentation.

based on an encoder-decoder network, where each layer of the decoder is used to detect objects of different scales, and the multi-scale fusion layer is used for semantic segmentation. DspNet [7] is another lightweight architecture for joint object detection and semantic segmentation. Its detection module is based on SSD [23], and the segmentation module is inspired by PSPNet [30]. More recently, TripleNet [8] and PairNet [8] are proposed. PairNet [8] uses a shared encoder-decoder structure. Each decoder layer is simultaneously used for detection and segmentation. During inference, only the last feature map is used for final segmentation. Compared to PairNet [8], TripleNet [8] goes a step further. It uses the attention skip-layer fusion to expand the feature map, the inner-connected module to increase the correlation between the two tasks, and the class-agnostic segmentation supervision to add a deep level of supervision.

III. Approach

A. Overview

Fig. 2 shows the overview architecture of our model. The model mainly consists of three modules: the encoder-decoder module, the object detection module, and the semantic segmentation module. The encoder-decoder module is composed of seven encoder layers and seven decoder layers, and each decoder layer is connected with the corresponding encoder layer by the "Skip Connection" network. The input image is processed by the encoder-decoder module, generating seven different sizes of feature maps in both encoder layers and decoder layers, which are subsequently used for the object detection and semantic segmentation. In the object detection module, the object detection is carried out by classifying the object candidates that are generated on the seven feature maps. In the semantic segmentation module, the feature maps are firstly upsampled to the same size (especially, the Multiple-Scale Atrous Convolution mechanism is applied on the last feature map), and then concatenated together to realize the segmentation. In the following, we will detail the encoder-decoder module, detection module, and the segmentation module.

B. Encoder-Decoder architecture

We use ResNet50 [32] as the backbone of the encoder network. ResNet50 is composed of five blocks, and each block outputs a feature map. To save the GPU memory usage, we use the last four feature maps of the ResNet50 and denote them as $E1$, $E2$, $E3$ and $E4$. In addition, we add three new residual layers after ResNet50, which generate three feature maps, which are denoted as $E5$, $E6$ and $E7$. The sizes of feature maps are gradually halved. If the resolution of the input image is $H \times W$, the resolution of $E7$ is $\frac{H}{256} \times \frac{W}{256}$.

The feature maps from the encoder covey relatively low-level semantic information. In order to improve the semantic information of the feature maps, we adopt the skip connection mechanism. The detail network structure of the skip connection is shown in the left corner of Fig. 2. The Squeeze-and-Excitation (SE) [13] module can obtain the importance of each feature channel, and increase the weight of useful features. We add SE [13] module to skip connection module, which is beneficial for improving the representation capabilities of decoder feature maps. Take an example to better understand the skip-connection mechanism, we use $E7$ and $E6$ to generate $D6$, to this end, we firstly upsample $E7$ and then concatenate it with $E6$ through SE module, then the concatenated feature map is processed by a series of convolutions to generate $D6$.

Through the skip connection, we generate decoder feature maps of different sizes with rich semantic information, which are denoted as $D6$, $D5$, $D4$, $D3$, $D2$ and $D1$. We use global average pooling on $D6$ to generate $D7$. The sizes of decoder feature maps are gradually doubled. Then the seven decoder feature maps are used for the object detection and semantic segmentation simultaneously.

C. Detection

To implement the object detection, we apply our self-defined PriorBox initialization mechanism to generate object candidates , based on which the detection is realized with the commonly-used classification and regression methods. The traffic scene is highly dynamic and complex, the appearance of objects are diverse (e.g., the aspect ratio of person is small while the aspect ratio of train is large). Therefore, our PriorBox initialization mechanism targets to generate candidates with various aspect ratios.

The scale of PriorBox for each feature map is computed as:
$if\ k >= 2$:

$$S_k = S_{min} + \frac{S_{max} - S_{min}}{m-2}(k-2), k \in [2, m] \tag{1}$$

$if\ k = 1$:

$$S_k = S_{smallest} \tag{2}$$

where m is the number of feature maps, k is the index of feature maps, and $S_{smallest}$, S_{min} and S_{max} are self-defined parameters. We set $m = 7$, $S_{smallest} = 0.04$, $S_{min} = 0.7$ and $S_{max} = 0.95$.

Inspired by some works like SSD [23], in the first five decoder layers (i.e., $D7$, $D6$, $D5$, $D4$, $D3$), we set 7 PriorBox with different ratios for each feature map location and the aspect ratios are $\alpha \in \{1,2,3,4,\frac{1}{2},\frac{1}{3},\frac{1}{4}\}$. In the last two layers of the decoder (i.e., $D2$, $D1$), we set 5 PrioBox with different ratios for each feature map location and the aspect ratios are $\alpha \in \{1,2,3,\frac{1}{2},\frac{1}{3}\}$. In all layers, for the aspect ratio of 1, we additionally add another PriorBox whose scale is $\sqrt{s_k s_{k+1}}$. For each PriorBox, the width is denoted as $w = S_k\sqrt{\alpha}$, the height is denoted as $h = S_k/\sqrt{\alpha}$. If the resolution of input image is 300×300, the scales for 7 decoder feature maps are $\{S_1 = 0.04$, $S_2 = 0.1$, $S_3 = 0.27$, $S_4 = 0.44$, $S_5 = 0.61$, $S_6 = 0.78$, $S_7 = 0.95\}$. Take an example, for each feature map location of $D1$, the $h \times w$ can be denoted as $\{12 \times 12$, 19×19, 17×8, 8×17, 21×7, 7×21 $\}$.

Our proposed PriorBox initialization mechanism can generate dense object candidates with special aspect ratios, which is beneficial for object detection in complex and dynamic traffic scenes.

D. Segmentation

Studies have proven that multi-scale feature fusion is useful for semantic segmentation [6]–[8], [30]. The main reason is that the feature maps in shallow layers tend to imply the local semantic information and the feature maps in deep layers tend to imply the global semantic information. Therefore, to get both the global and local semantic information, we upsample the feature maps of each decoder layer to the same resolution, and then concatenate them together to form a final feature map. Especially, in order to extract rich semantic information, we apply the MSAC mechanism in the last feature map. The Multi-Scale Atrous Convolution (MSAC) module is able to obtain the information with different receptive fields, allowing to effectively capture informative features. As shown in the Fig. 2, the MSAC network mainly consists of 5 branches, including a 1×1 convolution branch, 3 parallel 3×3 atrous convolution branches, and a global average pooling branch.

E. Loss Function

For the object detection, we use the similar loss functions that are widely adopted in the models of SSD [23] and Faster R-CNN [19]. For the semantic segmentation, the loss is the cross-entropy between predicted and target class distribution of pixels [27]. We use each feature map of the decoder to parse the labels of the semantic pixels separately. We upsample each segmentation logits to the same resolution as ground truth, use them to calculate the loss of each semantic segmentation, and accumulate all these losses. They can be regarded as a deep level of supervision in the entire learning process. Therefore we define the multi-task loss function as:

$$L = L_{det} + L_{seg} \tag{3}$$

$$L_{det} = L_{cls} + L_{reg} \tag{4}$$

$$L_{seg} = L_{infer\_fm} + L_{decoder\_fms} \tag{5}$$

Where L_{cls} is used to classify the object candidates, L_{reg} is used to refine the corresponding Priorbox, L_{infer_fm} means the cross-entropy between the fused feature map and ground truth, and $L_{decoder\_fms}$ means the cross-entropy between each decoder feature map and ground truth.

IV. Experiments

A. Setting

We perform experiments on the Cityscapes [14] dataset. Cityscapes is an image segmentation dataset collected in the traffic scenes, including 20,000 images with coarse annotations and 5000 images with high quality annotations. Following the setting in DspNet [7], 5000 images with high quality annotations are used in our experiments. Since the test dataset of Cityscapes does not provide detailed annotations, in actual experiments, we use training set (2975 images) for training and validation set (500 images) for testing.

For object detection, since the dataset does not provide the bounding box annotations, to make the dataset qualified for object detection, we compute four values (leftmost, rightmost, uppermost and nethermost) to form a bounding box containing a semantic segment. Following the DspNet [7], we set 8 classes for object detection and the rest as the background class.

For semantic segmentation, the original Cityscapes dataset contains pixel-level annotations for 33 classes. However, some classes are not important for the scene understanding or are

TABLE I

Comparison of mean average precision (mAP) results for each class on the Cityscapes-val dataset. The models are trained on Cityscapes-train. Their backbone is ResNet50 and input size is 300 × 300.

Method	person	rider	car	truck	bus	train	mbike	bike	**mAP**
DspNet [7]	23.0	27.5	52.8	30.8	48.1	40.5	19.8	25.1	33.4
BlitzNet [6]	28.7	31.8	63.9	34.1	57.2	45.1	20.6	**26.6**	38.5
PairNet [8]	21.6	28.8	48.8	33.2	53.4	49.3	14.2	22.4	34.0
TripleNet [8]	21.1	27.4	49.6	33.3	52.5	42.6	19.4	21.4	33.4
Our Method	**28.7**	**32.8**	**63.9**	**35.7**	**58.6**	**50.2**	**24.0**	26.5	**40.0**

TABLE II

Comparison of mean intersection over union (mIoU) results for each class on the Cityscapes-val dataset. The models are trained on Cityscapes-train. Their backbone is ResNet50 and input size is 300 × 300.

Method	road	swalk	build	wall	fence	pole	t.light	t.sign	veg.	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	**mIoU**
DspNet [7]	89.8	63.2	80.1	38.4	28.0	11.6	22.3	36.1	81.6	**49.2**	82.0	51.0	32.0	86.0	**63.2**	**71.0**	**62.2**	**36.8**	**51.0**	54.5
BlitzNet [6]	88.4	58.2	78.1	30.7	31.6	10.5	11.4	24.4	80.6	41.5	82.9	50.3	26.1	85.2	56.7	67.3	60.3	28.3	47.1	50.5
PairNet [8]	87.4	58.9	77.1	39.2	29.8	8.4	13.9	25.7	75.5	44.5	79.3	48.1	29.8	83.5	57.5	65.0	51.5	32.2	46.8	50.4
TripleNet [8]	87.7	60.6	77.7	38.3	30.1	9.1	12.0	29.5	80.7	45.8	80.5	49.1	27.8	84.8	63.0	68.9	49.9	30.5	48.4	51.3
Our Method	**90.7**	**65.0**	**81.0**	**45.3**	**33.6**	**17.8**	**26.4**	**38.5**	**83.2**	48.1	**83.8**	**53.5**	**33.1**	**86.4**	60.2	65.6	56.3	35.0	50.9	**55.5**

TABLE III

Joint detection and segmentation V.S. individual detection/segmentation. The models are trained on Cityscapes-train and tested on Cityscapes-val. Their backbone is ResNet50 and input size is 300 × 300.

Mode	**Det**	**Seg**	**mAP**	**mIoU**
Detection	√	–	36.3	–
Segmentation	–	√	–	55.0
Joint	√	√	**40.0**	**55.5**

rarely appeared in many scenarios, following the setting in [33], [34], we use 19 classes for semantic segmentation, and the rest as the background class.

B. Metrics

For object detection, mean average precision (mAP) is generally used to evaluate the performance of object detection. For semantic segmentation, mean intersection over union (mIoU) is generally used to evaluate the performance of semantic segmentation. Therefore, we take the mAP as the metric for object detection and mIoU as the metric for semantic segmentation.

C. Implementation details

Our proposed method is coded in python 3.6.9 and pytorch 1.2. All experiments run on a single NVIDIA GeForce RTX 2080Ti GPU. In all our experiments, we use SGD [35] to optimize the network, with a batch size of 6 images, and the input size of each image is resized to 300 × 300. The total number of epoch in the training stage is 320, where the initial learning rate is set to 0.0005. The learning rate decreases by a factor 2 at epoch 80/160/240, respectively.

D. Experimental design and results analysis

Comparison with baseline methods in object detection. Table I shows the performance of baseline methods and our method in object detection. Our method exhibits best performance on the detection of person, rider, car, truck, bus, train and mbike, and achieves 3.9% accuracy improvement compared with the second best model BlitzNet [6] and 19.8% improvement compared with the recently-proposed TripleNet [8]. Fig. 3 shows some samples of object detection results, from which we can observe that our model is effective. The reasons are three-fold: 1) we adopt the skip-connection mechanism and involve the SE module in the skip-layer connection, so the feature maps covey the informative global and local information; 2) our self-defined PriorBox initialization mechanism generates denser and multi-scale object candidates, which contributes to detecting objects with special aspect ratios; 3) the detection branch shares the same feature maps with the semantic segmentation branch, which benefits the object detection to use the semantic segmentation information.

Comparison with baseline methods in semantic segmentation. Table II describes the performance of baseline methods in semantic segmentation. Our method achieves 1.8% accuracy improvement compared with the second best model DspNet [7] and 8.2% improvement compared with the recently-proposed TripleNet [8], and exhibits the best performance on most semantic classes. Fig. 3 shows some samples of semantic segmentation results, from which we can observe that our model is effective. The skip connection in the encoder-decoder network is significant for the good performance. In addition, the MSAC module also contributes to the good performance since it is able to extract the highly abstract feature map by utilizing multi-scale atrous convolutions.

Joint detection and segmentation V.S. individual detection/segmentation. As shown in Table III, compared to indi-

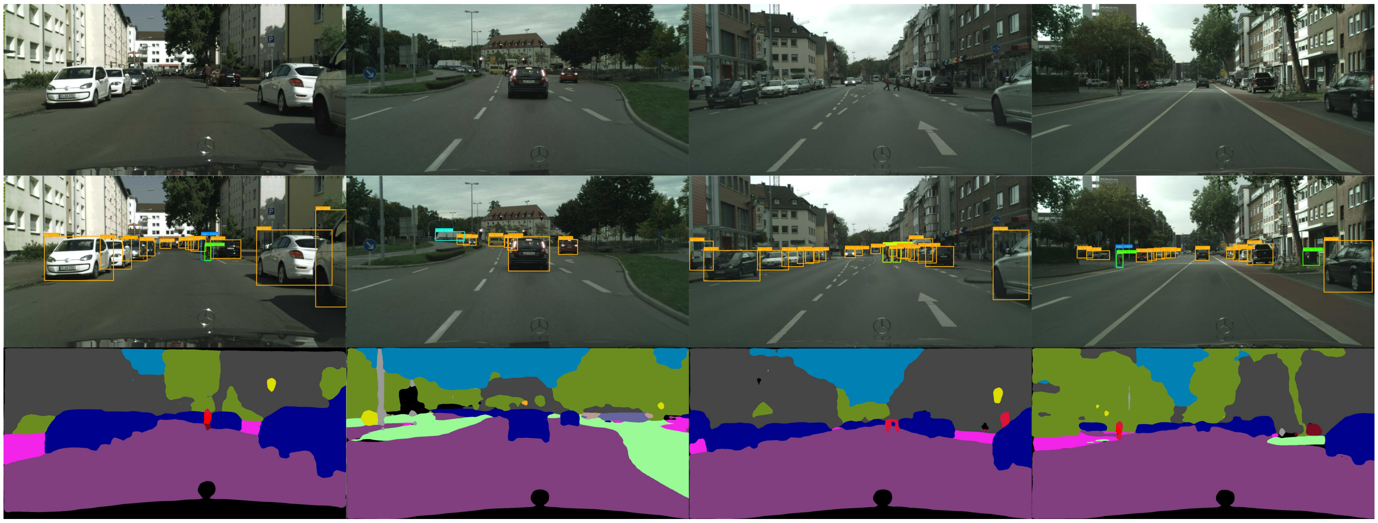

Fig. 3. Visualization experimental results of our method on Cityscapes-val dataset. Specifically, the first row shows the input images, the second row shows our detection results, and the last row shows our segmentation results.

TABLE IV
ABLATION EXPERIMENTS OF OUR MODEL. THE MODELS ARE TRAINED ON CITYSCAPES-TRAIN AND TESTED ON CITYSCAPES-VAL. THEIR BACKBONE IS RESNET50 AND INPUT SIZE IS 300 × 300. "BASE" MEANS USE DECONVOLUTION TO ENLARGE FEATURE MAPS, "+SKIP" MEANS USE SKIP CONNECTION TO ENLARGE FEATURE MAPS, "+SKIP+MSAC" MEANS BOTH SKIP CONNECTION AND MSAC ARE ADOPTED.

Configuration	mAP	mIoU
Base	13.8	38.0
Base+Skip	38.8	54.4
Base+Skip+MSAC	**40.0**	**55.5**

vidual detection or segmentation network, the joint detection and segmentation network exhibits better performance. The mAP is improved from 36.3% to 40.0% and the mIoU is improved from 55.0% to 55.5%, which demonstrates that the two tasks are mutually beneficial. On the one hand, object detection can be used as prior knowledge to assist semantic segmentation, on the other hand, semantic segmentation can provide context information and semantic features for object detection. From the perspective of learning, the branches of object detection and semantic segmentation share the same feature map extraction network, the backward procedure updates the parameters of the encoder-decoder network, guiding the network to learn the features that simultaneously benefit the detection task and segmentation task.

Effectiveness of MSAC module and skip connection. Table IV shows the experiment results of our model with different network configurations. When skip connection and MSAC are not used, the accuracies of detection and segmentation are respectively 0.138 and 0.380. When the skip connection module is added, the accuracies of detection and segmentation are improved to 0.388 and 0.544, respectively. When both skip connection and MSAC are added, the accuracies of detection and segmentation are improved to 0.400 and 0.555, respectively. Experiments prove that it is effective to extract the feature maps that fuse the information in both shallow layers and deep layers by employing the skip-connection, and it is significant for semantic segmentation to enlarge the receptive field through MSAC mechanism.

E. Extention experiment and results analysis on VOC2012-Segmentation dataset

In this subsection, to demonstrate the generalization ability and robustness of our proposed model, we perform experiments on the PASCAL VOC [15] dataset. The PASCAL VOC dataset consists of VOC2007 and VOC2012. The VOC2007 and VOC2012 datasets are often used to evaluate the performance of object detection and semantic segmentation. The number of classes for object detection and semantic segmentation is 20. For the dataset containing the semantic segmentation label, the VOC2007-segmentation dataset consists of training set (206 images), validation set (213 images), test set (210 images), the VOC2012-segmentation dataset consists of training set (1464 images), validation set (1449 images), without the test set. We use the VOC2012-segmentation dataset for experiments, using training set (1464 images) for training and validation set (1449 images) for testing.

Table V and Table VI show the performance of baseline methods and our method in object detection and semantic segmentation, from which we can observe that our model outperform other methods on most object classes and segmentation types, and achieves the highest overall performance on both object detection task and semantic segmentation task, surpassing the second best model BlitzNet [6] by 2.5% mAP in object detection and TripleNet [8] by 1.4% mIoU in semantic segmentation. Fig. 4 shows some samples of joint object detection and semantic segmentation of our method, and we can observe that our model exhibits good performance. Experimental results prove that our model is robust on different kinds of datasets.

TABLE V

COMPARISON OF MEAN AVERAGE PRECISION (MAP) RESULTS FOR EACH CLASS ON THE VOC2012-SEGMENTATION-VAL DATASET. THE MODELS ARE TRAINED ON VOC2012-SEGMENTATION-TRAIN. THEIR BACKBONE IS RESNET50 AND INPUT SIZE IS 300 × 300.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**mAP**
DspNet [7]	77.6	67.3	68.8	56.7	33.8	81.6	53.9	86.5	35.1	66.2	52.3	76.4	71.6	81.6	68.8	34.9	66.4	58.0	81.4	69.5	64.4
BlitzNet [6]	81.1	75.3	73.8	63.1	38.0	87.0	**61.3**	85.8	40.7	72.5	53.4	74.4	72.8	84.5	**74.0**	41.2	68.5	57.6	**86.7**	**76.4**	68.4
PairNet [8]	81.6	69.1	73.5	59.8	36.5	82.9	57.1	82.5	31.8	71.7	53.7	76.2	73.8	**84.7**	70.1	39.0	68.8	61.2	86.4	72.2	66.6
TripleNet [8]	**84.2**	74.7	**81.6**	61.2	34.9	82.9	55.0	**86.6**	32.3	72.4	54.0	**81.8**	73.5	82.9	67.3	31.2	70.2	59.7	84.7	73.9	67.2
Our Method	81.9	**75.3**	75.7	**68.2**	**42.5**	**88.5**	61.2	84.0	**42.3**	**74.8**	**57.0**	79.6	**76.2**	83.4	71.3	**41.4**	**72.5**	**63.5**	86.0	76.3	**70.1**

TABLE VI

COMPARISON OF MEAN INTERSECTION OVER UNION (MIOU) RESULTS FOR EACH CLASS ON THE VOC2012-SEGMENTATION-VAL DATASET. THE MODELS ARE TRAINED ON VOC2012-SEGMENTATION-TRAIN. THEIR BACKBONE IS RESNET50 AND INPUT SIZE IS 300 × 300.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**mIoU**
DspNet [7]	73.3	**35.2**	72.0	57.8	62.3	**84.9**	**75.7**	75.2	22.6	64.4	50.3	68.4	61.9	**68.6**	70.6	**50.3**	63.1	36.7	76.6	**61.8**	61.6
BlitzNet [6]	74.1	31.3	73.4	54.3	59.6	82.4	73.3	73.0	20.5	64.9	46.0	66.7	61.1	65.3	67.6	47.4	57.3	37.1	76.3	58.9	59.5
PairNet [8]	69.5	31.3	66.4	55.6	56.6	78.3	70.7	68.1	24.3	58.3	44.4	63.0	63.0	60.9	67.4	44.3	59.1	36.1	73.8	58.2	57.5
TripleNet [8]	74.5	30.8	74.8	57.4	**65.5**	83.8	72.5	77.2	24.0	68.1	47.9	70.8	64.8	67.1	70.0	49.5	**69.4**	**41.5**	76.9	61.6	62.4
Our Method	**76.8**	34.2	**77.3**	**61.6**	61.1	84.3	73.7	**78.2**	**26.1**	**72.2**	**50.8**	**73.8**	**70.1**	65.4	**71.2**	44.4	66.3	38.5	**78.2**	61.2	**63.3**

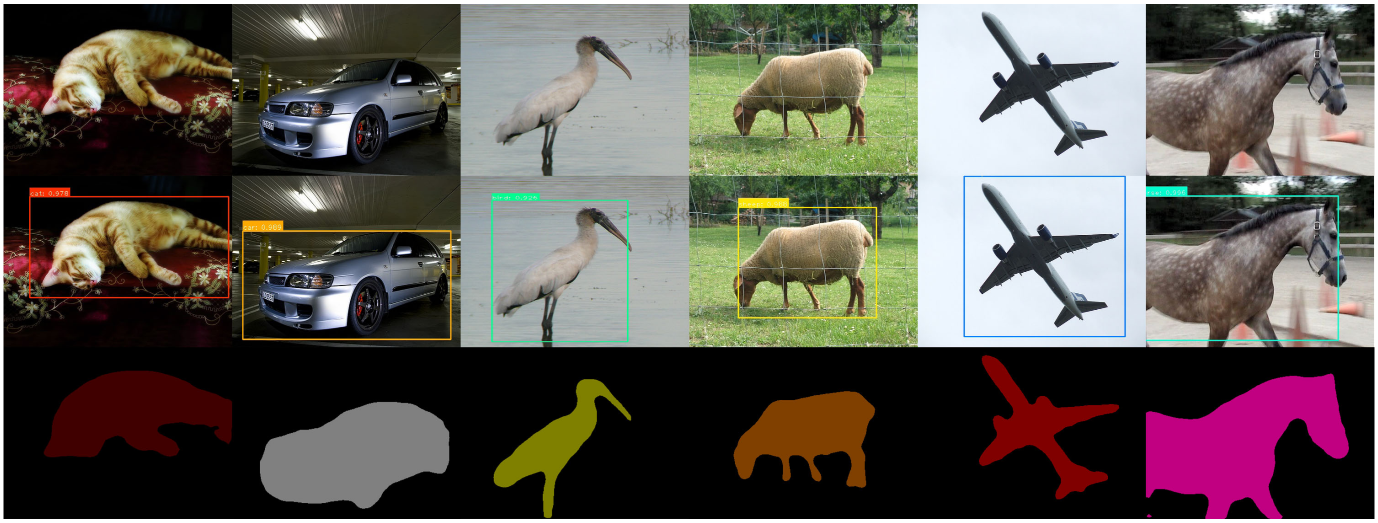

Fig. 4. Visualization experimental results of our method on VOC2012-segmentation-val dataset. Specifically, the first row shows the imput images, the second row shows our detection results, and the last row shows our segmentation results.

Our method achieves better results on the VOC2012-segmentation dataset than that on the Cityscapes dataset. One main reason is that the Cityscapes dataset is collected in the dynamic and complex traffic scenes, with various backgrounds and diverse small-scale objects. In the contrast, the images in the VOC2012-segmentation dataset often contain only one or a few objects and the backgrounds of images are not complex as that in the Cityscapes dataset.

V. CONCLUSION

In this paper, we propose a deep model for joint object detection and semantic segmentation in traffic scenes. We involve the SE network in the skip-connection mechanism to form the informative feature maps, propose a new PriorBox initialization mechanism to generate denser candidate objects, and adopt MSAC to enlarge the receptive field and explore global and local semantic information. Our model achieves the competitive results on a variety of experiments, from which we draw some conclusions: 1) object task and segmentation are mutually beneficial, 2) the proper PriorBox initialization mechanism is important for the object detection, for example, the traffic scenes contain various object with diverse appearance, thus it is important to set PriorBox with different aspect ratios, and 3) the skip-connection mechanism and MSAC are significant for extracting the informative global and local semantic feature maps, which are crucial to improve overall performance. In the future, we plan to further strengthen the correlation between object detection and semantic segmentation so that they can achieve better mutual benefit.

References

[1] D. Wang, C. Devin, Q.-Z. Cai, F. Yu, and T. Darrell, "Deep object-centric policies for autonomous driving," *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8853–8859, 2018.

[2] K. Yang, L. M. Bergasa, E. Romera, R. Cheng, T. Chen, and K. Wang, "Unifying terrain awareness through real-time semantic segmentation," *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1033–1038, 2018.

[3] S. Fidler, R. Mottaghi, A. L. Yuille, and R. Urtasun, "Bottom-up segmentation for top-down detection," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3294–3301, 2013.

[4] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. L. Yuille, "The role of context for object detection and semantic segmentation in the wild," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 891–898, 2014.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[6] N. Dvornik, K. Shmelkov, J. Mairal, and C. Schmid, "Blitznet: A real-time deep network for scene understanding," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4174–4182, 2017.

[7] L. Chen, Z. Yang, J. Ma, and Z. Luo, "Driving scene perception network: Real-time joint detection, depth estimation and semantic segmentation," *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1283–1291, 2018.

[8] J. Cao, Y. Pang, and X. Li, "Triply supervised decoder networks for joint detection and segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 7392–7401, 2019.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *CoRR*, vol. abs/1412.7062, 2014.

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, pp. 834–848, 2016.

[11] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *ArXiv*, vol. abs/1706.05587, 2017.

[12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

[13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2017.

[14] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016.

[15] M. Everingham, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2009.

[16] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2013.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, pp. 1904–1916, 2014.

[18] R. B. Girshick, "Fast r-cnn," *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, 2015.

[19] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2015.

[20] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6154–6162, 2017.

[21] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask r-cnn," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.

[22] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2015.

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[24] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd : Deconvolutional single shot detector," *ArXiv*, vol. abs/1701.06659, 2017.

[25] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017.

[26] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, 2016.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, 2015.

[28] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 2481–2495, 2017.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[30] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230–6239, 2016.

[31] I. Kokkinos, "Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5454–5463, 2016.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

[33] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," *ArXiv*, vol. abs/1811.11721, 2018.

[34] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3684–3692, 2018.

[35] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, pp. 177–186, Springer, 2010.