SA-YOLOv3: An Efficient and Accurate Object Detector Using Self-attention Mechanism for Autonomous Driving

Daxin Tian, *Senior Member, IEEE,* Chunmian Lin, Jianshan Zhou, Xuting Duan, *Member, IEEE,*
Yue Cao, Dezong, Zhao, *Senior Member, IEEE,* and Dongpu Cao

***Abstract*—Object detection is becoming increasingly significant for autonomous-driving system. However, poor accuracy or low inference performance limits current object detectors in applying to autonomous driving. In this work, a fast and accurate object detector termed as SA-YOLOv3, is proposed by introducing dilated convolution and self-attention module (SAM) into the architecture of YOLOv3. Furthermore, loss function based on GIoU and focal loss is reconstructed to further optimize detection performance. With an input size of 512×512, our proposed SA-YOLOv3 improves YOLOv3 by 2.58 mAP and 2.63 mAP on KITTI and BDD100K benchmarks, with real-time inference (more than 40 FPS). When compared with other state-of-the-art detectors, it reports better trade-off in terms of detection accuracy and speed, indicating the suitability for autonomous-driving application. To our best knowledge, it is the first method that incorporates YOLOv3 with attention mechanism, and we expect this work would guide for autonomous-driving research in the future.**

***Index Terms*—Autonomous driving, object detection, attention mechanism, deep learning, YOLOv3, intelligent transportation systems.**

I. Introduction

As an essential part of autonomous driving, environmental perception must have the capability of fast and accurate object detection in real-world condition, in order to ensure safe and correct driving behavior and decision [1-2]. In other words, an object detector applicable for autonomous driving should satisfy the following two prerequisites. First, highly accurate and robust detection performance is required, and therefore detector could accurately handle multiple object localization and recognition in the complex and various traffic scene. Second, fast inference is also important for the real-time response and low latency of autonomous-driving system. Consequently, autonomous-driving applicable object detector needs to achieve trade-off between high accuracy and fast efficiency.

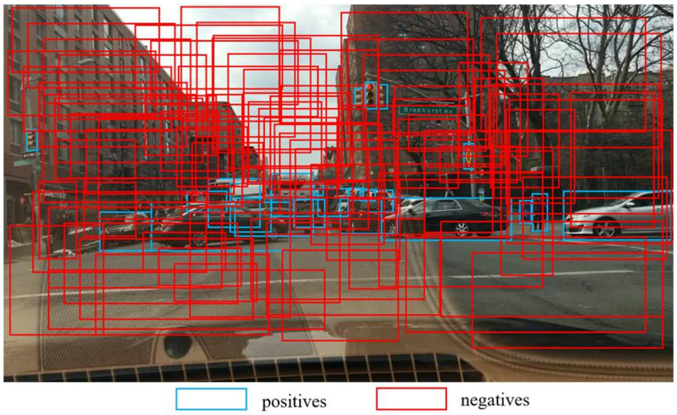


Fig. 1. The extreme positive-negative imbalance in one-stage detector. Postive sample is labeled by blue line, while bounding box with red line presents negative example.

With the introduction of deep learning, object detection has experienced remarkable development and progress, and much more promising performance has been further reported by adopting advanced convolutional architecture. Currently, CNN-based detectors can be roughly categorized into two classes: two-stage and one-stage object detector. To be specific, two-stage detector, e.g., R-CNN (region convolutional neural network) model [3-5], generally reports accurate detection performance by the guidance of region proposal and bounding-box (bbox) refinement. However, slow inference speed hinders its application in real-time system due to heavy computational cost in region proposal generation. In contrast, one-stage detector, i.e., YOLO model [6-8], has extremely fast detection speed in a single inference that formulates object detection as a simple regression problem which refers to simultaneous computation of bounding-box regression and classification on convolutional feature map. Nevertheless, without the help of region proposal, one-stage method easily suffers from poor localization or classification result, thereby reporting much inaccurate performance than that of two-stage detector. Generally, existing object detectors fail in either inefficiency or inaccuracy performance, and consequently, have a limitation in application to autonomous driving.

To investigate a fast but accurate object detector suitable for autonomous-driving system, in this work we mainly focus on one-stage method and attempt to explore its potentials for performance improvement. As mentioned above, one-stage

Corresponding anthor: Xuting Duan (duanxuting@buaa.edu.cn).

D. Tian, C. Lin, J. Zhou, X. Duan, Y. Cao are with Beijing Advanced Innovation Center for Big Data and Brain Computing, Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, School of Transportation Science and Engineering, Beihang University, Beijing 100191, China.

D. Zhao is with the Department of Aeronautical and Automotive Engineering, Loughborough University, Loughborough, LE11 3TU, United Kingdom.

D. Cao is with the Department of Mechanical and Mechatronics Engineering, University of Waterloo, 200 University Ave West, Waterloo ON, N2L3G1 Canada

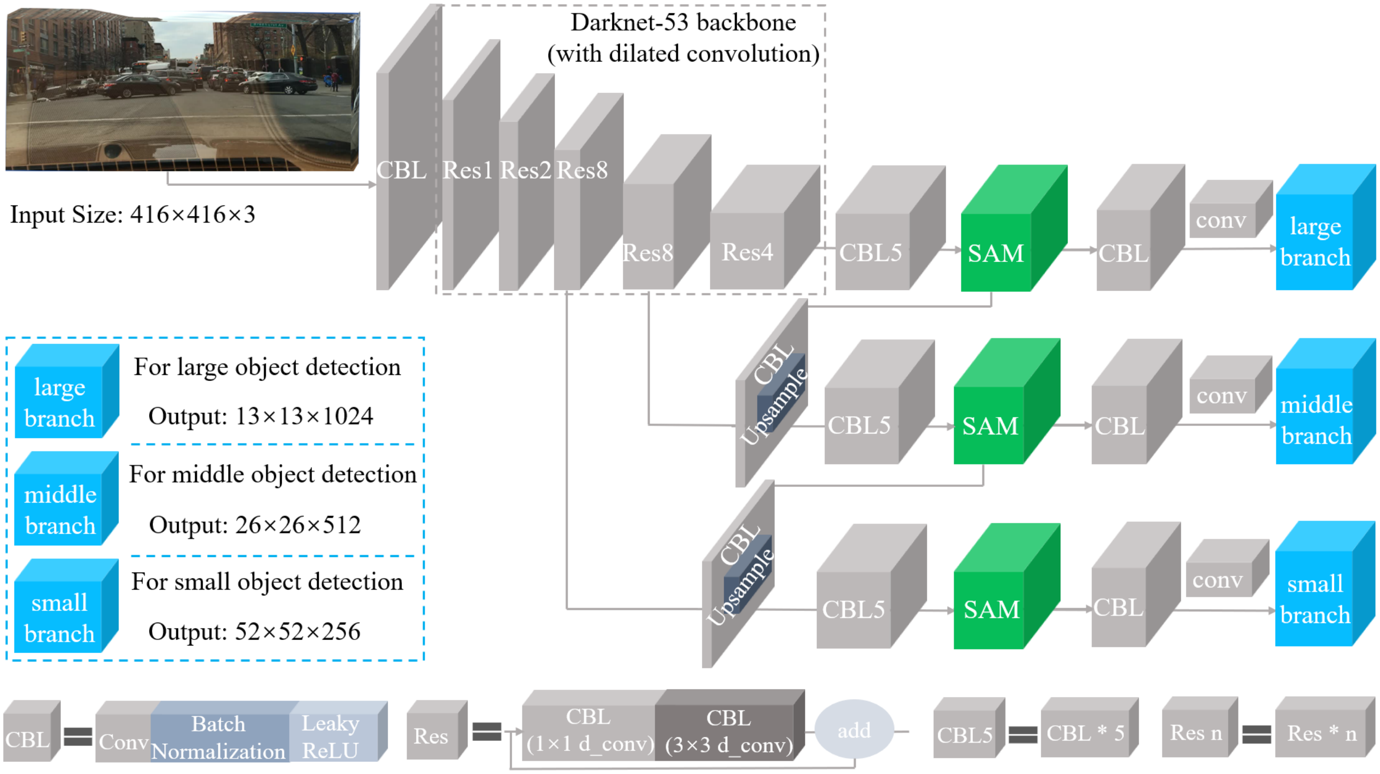


Fig. 2. Overview of architecture of SA-YOLOv3. Compared with YOLOv3, SA-YOLOv3 replaces standard convolution with dilated convolution in Darknet-53 backbone, and simultaneously inserts self-attention module (SAM) into the detection head. Moreover, large, middle and small branches are used for different sizes of object detection, as illustrated in the left-bottom side. And the bottom describe the component of CBL, Res, CBL5 and Res*n blocks.

detector is generally less accurate than two-stage method, and there are three main reasons as follows. Firstly, the ubiquitous problem of extreme class imbalance in one-stage detector damages detection accuracy a lot, as described in Fig. 1. To solve this problem, RetinaNet [9] with focal loss is proposed to eliminate the effect of foreground-background imbalance problem, but it hardly eradicates its imbalance problem. More importantly, discriminative feature and their relation are difficult to discover and capture in a single inference, and consequently, irrelevant or negative feature information causes inaccurate box regression and mislocalization. This is an essential problem resulting in poor detection performance, and we aim to improve one-stage detection performance from this perspective.

Inspired by human vision mechanism, a fast and accurate object detector termed as self-attention YOLOv3 (SA-YOLOv3) is proposed in this work. Based on efficient YOLOv3 [8], SA-YOLOv3 is designed by incorporating dilated convolution [10] with self-attention module (SAM). On one hand, dilation convolution can enlarge receptive field where convolutional filter can adaptively focus on contextual information and preserve discriminative feature without losing resolution. On the other hand, self-attention module (SAM) is introduced to capture global feature relation and learn the importance of feature at different positions in spatial space. It can suppress irrelevant information by using global dependency, and simultaneously highlight useful feature regions to guide for accurate detection. To further boost detection performance, loss function based on GIoU [11] and focal loss is reconstructed for alleviating class imbalance problem and improving box localization performance. Extensive evaluation experiments on KITTI [12] and BDD100K [13] benchmarks are conducted to validate the effectiveness of proposed algorithm. Compared with YOLOv3 and other advanced detectors, SA-YOLOv3 achieve competitive trade-off between detection accuracy and speed, indicating the suitability for autonomous driving. In conclusion, our contributions are mainly summarized as follows:

1) Self-attention YOLO-v3 (SA-YOLOv3) is proposed by introducing dilated convolution and self-attention module (SAM) into the architecture of YOLOv3. To our best knowledge, it is the first object detector that incorporates YOLOv3 with attention mechanism.

2) Based on GIoU and focal loss, targeted loss function is reconstructed to alleviate class imbalance problem in one-stage detector, pursuing for better detection performance.

3) Comprehensive experiments are performed on KITTI and BDD100K benchmarks, and evaluation results demonstrate that SA-YOLOv3 outperforms YOLOv3 by a large margin both on detection accuracy and speed, and also reports better balance in terms of detection accuracy and speed when compared to other state-of-the-art detectors.

The remainder of this paper is organized as follows: related works are summarized in section II; the proposed self-attention YOLOv3 and experimental analysis are demonstrated in section III and section IV, respectively. And section V concludes our contributions in this paper and discusses future works.

II. Related Works

A. Object Detection

Object detection is always an important issue in the field of computer vision and autonomous driving, which can be formulated as object localization and recognition in an image. Traditional object detection is mainly image processing- or feature-based algorithms [14-17], however, hand-crafted features cannot adaptively address various and complex real-world scenes. It therefore reports unsatisfactory detection performance. Recent years have witnessed the promising development and progress with CNN-based object detector, and as mentioned above, it can be divided into two categories: two-stage and one-stage detector.

1) Two-stage detector: R-CNN [3] is the most representative two-stage object detector that firstly builds region proposal generation followed by object classification and box regression for localization in the second stage. Moreover, Fast R-CNN [4] and Faster R-CNN [5] are further proposed for architecture optimization and performance improvement. Apart from R-CNN detector, many two-stage object detectors are developed for different purposes. To explore better detection performance, Li et al. [18] optimizes region-based method with fully convolutional architecture, and position-sensitive score map is adopted for significant performance gains. To investigate the effectiveness of multi-scale feature representation, Lin et al. [19] builds bottom-up and top-down architecture to fuse multi-level semantic features, and consequently, more accurate performance is observed with rich feature representations. Furthermore, to address the scale problem in real-world object detection, Hu, et al. [20] proposes SINet with scale-insensitive module, which incorporates context-aware RoI pooling and multi-branch decision network for multi-scale object detection. It therefore can accurately detect object of various sizes and scales, and achieve promising performance in vehicle detection. To speed up the inference in two-stage detector, MS-CNN [21] is designed by using intermediate convolutional architecture, and the proposed sub-network accelerates two-stage detection pipeline. Generally, two-stage detector can achieve on par with or even better accuracy performance than that of human; however, multi-stage inference incurs extensive computation, and therefore hinders its application in real-time system.

2) One-stage detector: The most typical one-stage detector is YOLO [6] that directly predicts bounding-box localization and class possibility score on convolutional feature map in a single inference. And later, YOLOv2 [7] is proposed by introducing batch normalization [22] and multi-scale anchor scheme, to address inaccurate detection problem in YOLO. Recently, YOLOv3 [8] is further enhanced by more promising backbone network and multi-branch architecture for detecting object of various sizes. To be specific, Darknet-53 backbone with residual blocks [23] is designed for multi-scale feature extraction, followed by three branch architectures for large, middle and small object detection. Besides, there are also numerous researches on one-stage detector, and they are mostly dedicated to optimize detection accuracy [9, 24-30]. Liu et al. [24] designs single-shot multi-box detector (SSD) that introduces default prior and combines multi-scale feature map from different layers to capture richer feature representations. Intuitively, it can easily handle object of various sizes and greatly improve detection accuracy. Based on the architecture of SSD, several variants are subsequently developed including FSSD [25], DSOD [26], RefineDet [27] and STOD [28]. These methods only achieve marginal accuracy improvement but at the cost of significant computational efficiency, which is unsuitable for real-time application. Liu et. al [29] filstly adopts dilation convolution for object detection, which is used to enlarge receptive field and capture discriminative feature map. It is referred to as RFBNet that reports more accurate and robust detection performance. Furthermore, small object detection is an intractable problem for one-stage detector. Zhao et al. [30] proposes comprehensive feature enhancement network (CFENet) by combining feature fusion block (FFB) and comprehensive feature enhancement (CFE). FFB is responsible for concatenating multi-scale feature from different layers, and CFE is designed to further enhance shallow feature across different layers. While great performance gains in small object illustrates the effectiveness of CFENet and its components, complex architecture significantly increases the inference time, thus limiting the application to autonomous driving .

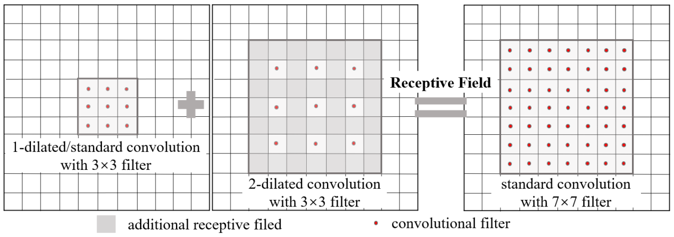
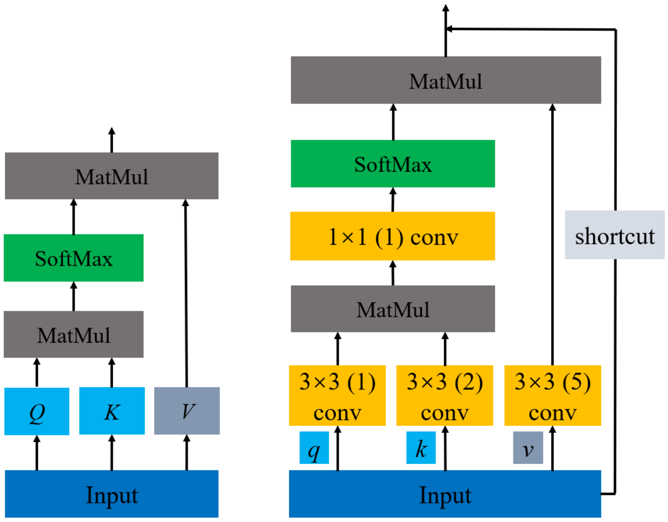

Receptive Field
1-dilated/standard convolution with 3×3 filter
2-dilated convolution with 3×3 filter
standard convolution with 7×7 filter
additional receptive filed
convolutional filter

Fig. 3. Overview of dilated convolution. A 1-dilated 3 × 3 convolution (left) followed by a 2-dilated 3 × 3 convolution (middle) have a larger receptive field on par with that of 7 × 7 standard convolution (right), under the similar computation budget.

B. Attention Mechanism

The concept of attention mechanism is early derived from natural language processing (NLP), and is successfully applied for machine translation [31-33]. Gradually, it extends in many fields, e.g. audio recognition [34], and its applications in computer vision [35-37] can be mainly classified into the following two folds: visual attention and self-attention mechanism.

1) Visual attention mechanism: Visual attention [38] in recurrent architecture is used to extract information from image or video by adaptively selecting a sequence of local region or location. In terms of object detection, one representative application of visual attention is RelationNet [39], which exploits object geometry and appearance information to model feature relation and capture attentive feature between region proposals. However, RelationNet is designed for two-stage detection pipeline, and its complexity is difficult to deploy in practice. Moreover, dilation convolution can be also considered as an application of attention mechanism in



(a) Transformer (b) Self-attention Module (SAM)

Fig. 4. Overview of architecture of self-attention module (SAM). (a) The attention unit in Transformer. (b) The self-attention module in this work. It is noted that we introduce 1-, 2- and 5-dilation convolution into SAM.

CNN. To be specific, dilated rate is introduced to control the size of receptive field, which allows filter to focus on contextual feature and its relation in a local region. More importantly, dilated convolution is a lightweight operation and can be efficiently plugged into other architecture without extra computational budget. The effectiveness, efficiency and superiority in object detection has been validated by RFBNet [29].

2) Self-attention mechanism: Inspired from self-attention mechanism in Transformer [32], recently non-local operation [40] is proposed to capture long-range dependencies from an image or video. It is described as a generic block to learn global feature relation by computing the response at a position as a weighted sum of feature at all positions. However, global pixel-wise computation incurs extensive computational operations and thus influences inference speed heavily. Besides, Yi et al. [41] introduces attentive single-shot detector (ASSD) by modeling feature relation in the spatial space. With reliable guidance of global feature information, ASSD achieves considerable performance improvement in accuracy. However, these methods usually neglect the significant effect of local contextual information in object detection, thereby reporting unsatisfactory detection accuracy.

III. Self-attention YOLOv3

In this work, we consider local semantic feature and global feature relation simultaneously, and design an efficient and accurate object detector based on YOLOv3. As mentioned above, YOLOv3 [8] is one of the state-of-the-art one-stage detectors, and its architecture can be mainly divided into two parts: Darknet-53 backbone and multi-branch detection head network. Darknet-53 mainly consists of twenty-three residual blocks, each of which uses successive 1×1 and 3×3 convolution with doubly increasing filter channels, as well as shortcut connection between input and convolutional output. In detection head network, it predicts boxes at 3 different scales by using similar concept to feature pyramid network [19]. To be specific, taking the input with a resolution of $N \times N$ (e.g. N=416) for example, the size of feature map in three-scale branches is 13×13, 26×26 and 52×52 by a reduction size of 32, 16 and 8, respectively. And upsampled feature map is further concatenated with earlier feature map to merge pyramid feature, which allows the model to capture multi-scale semantic and fine-grained feature information from shallow layers. However, there are two main problems of YOLOv3 that potentially damage its detection accuracy. On one hand, fixed and limited reception field in Darknet-53 architecture makes it hard to learn discriminative feature. On the other hand, contextual information and global relation would be gradually reduced with the increment of feature layers, the result of which may lower model performance a lot. To address these problems, we design self-attention YOLOv3 (SA-YOLOv3) that incorporates dilated convolution with self-attention module (SAM) into the architecture of YOLOv3, as illustrated in Fig. 2.

A. Dilation Convolution

To learn discriminative feature map, we replace standard convolution with dilated convolution in Darknet-53 backbone. Dilation convolution has been referred as convolution with a dilated filter in the past, and plays a key role in the átrous algorithm in wavelet decomposition [42-43]. Later, it is further adopted semantic segmentation [10] to aggregating multi-scale contextual feature maps without losing resolution of image. Mathematically, convolutional operation (*) between two functions can be described as follows (1):

$$(f * g)(r) = \sum_{m+n=r} f(m) * g(n) \tag{1}$$

where f is a discrete function with kernel size of m, g is the filter with the size of n, and r indicates the size of receptive field. Besides, dilated convolution is formulated as (2): where k denotes dilated rate.

$$(f *_k g)(r) = \sum_{m+kn=r} f(m) * g(n) \tag{2}$$

It is noted that dilation convolution is based on convolution operation in our work, and cannot make great effort to reconstruct dilated filter. By using multiple dilated rates, it can apply the same filter at various ranges to receive different receptive field. According to hybrid dilated convolution scheme [44], we place 1- and 2-dilated rate for 1×1 and 3×3 residual convolution in Darknet-53, respectively. Specifically, 1-dilated convolution is equal to the standard convolution in terms of receptive field, and a 1-dilated 3×3 convolution followed by 2-dilated 3×3 convolution has a receptive field of 7×7, as illustrated in Fig. 3. Consequently, there are at least two key advantages of adopting dilated convolution. First, it provides larger receptive field, and allows the model to focus on local feature information. Second, the introduction of dilated convolution rarely incurs additional computational cost, and thus ensures fast convolutional operation and efficient inference.

TABLE I
ANCHOR BOX RESULST BY K-MEANS CLUSTER FOR KITTI AND BDD100K TRAINING SET

		Anchor0	Anchor1	Anchor2
KITTI	Small	$(13, 30)$	$(23, 53)$	$(17, 102)$
	Middle	$(45, 76)$	$(27, 172)$	$(67, 116)$
	Large	$(49, 240)$	$(82, 170)$	$(118, 200)$
BDD100K	Small	$(7, 10)$	$(14, 24)$	$(27, 43)$
	Middle	$(32, 97)$	$(57, 64)$	$(92, 109)$
	Large	$(73, 175)$	$(141, 178)$	$(144, 291)$

B. Self-attention Module (SAM)

To highlight useful region and model feature relation, we further study self-attention mechanism that is successfully adopted by Transformer [32] in machine translation. In Transformer, input is divided into three compponents of query *(Q)*, key *(K)* and value *(V)*, as shown in Fig. 4 (a). Dot products of query with all keys are firstly computed, and then softmax function is placed on matrix multiplication result to obtain its weight on the value. It is termed as self-attention mechanism where long-range dependencies between contextual sentence can be considered and captured.

In this work, we design self-attention module (SAM) by the motivation of Transformer, and thus formulate it as sole attention mechanism to model feature relation. As demonstrated in Fig. 4 (b), the input is firstly divided into *q*, *k* and *v*, and 3×3 convolution with 1-, 2- and 5-dilated rate followed by softmax activation is performed. Before the final shortcut connection, 1×1 convolution plays the role of bottleneck for parameter reduction. With the help of dilated convolution, SAM can capture global feature relation, and simultaneously focus on local semantic information.

Mathematically, our proposed SAM can be expressed as follows in (3) and (4): we firstly formulate *q* and *k* feature maps by linearly scaling the input, and *v* maintains the same size with input. Here, x_s is used to denote feature map of input x under scale factor *s*, which indicates the number of filter and channel are reduced *s* times than that of input simultaneously; in this work *s* is set to 2 for model simplicity and efficient computation. And W_{qs}^T and W_{ks}^T standard the weight matrix of convolutional operation to obtain self-attention feature maps $q(x_s)$ and $k(x_s)$, respectively.

$$q(x_s) = W_{qs}^T x_s \tag{3}$$

$$k(x_s) = W_{ks}^T x_s \tag{4}$$

In (5), matrix multiplication between *q*, *k* and *v* feature maps is computed, followed by softmax function.

$$att(q, k, v) = softmax(q(x_s)k(x_s)^T)v(x) \tag{5}$$

In (6), *y* is finally outputted by adding attentive feature map with the input *x*.

$$y = x + att(q, k, v) \tag{6}$$

As illustrated in Fig. 2, SAM is placed between five CBL blocks and final two convolutional layers. Additionally, we make a slight improvement for five CBL blocks to further reduce computation cost: each 1×1 filter is used as bottleneck by doubling its channels, and accordingly the number of 3×3 filter is halved. Therefore, our proposed SA-YOLOv3 only requires 9.88×10^{10} FLOPs with an input size of 512×512, compared to 9.92×10^{10} FLOPs of YOLOv3. Generally, there are also two significant merits of SAM. Compared with extensive operation in non-local network [40], SAM is lightweight and it can be efficiently plugged into other architectures with rarely negligible computational cost. More importantly, it can model global feature relation and meanwhile capture local contextual information, which is superior to the attentive mechanism in ASSD [41].

C. Reconstruction of loss function

The loss function in YOLOv3 consists of squared sum error (SSE) and binary cross-entropy (BCE) to measure box localization and classification performance. However, due to heavy foreground-background imbalance, training loss is dominated by numerous negative or easy samples, thus leading to significant mislocalization and classification problems. In this work, we aim to reconstruct more targeted objective function for better convergence and performance in model training. To enhance box localization performance, we formulate generalized intersection-of-union (GIoU) [11] metric as localization loss L_{loc}. Mathematically, IoU, GIoU and L_{loc} can be described as (7)-(9):

$$IoU = \left| \frac{B_g \cap B_p}{B_g \cup B_p} \right| \tag{7}$$

$$GIoU = IoU - \left| \frac{C_{gp} - B_g \cup B_p}{C_{gp}} \right| \tag{8}$$

$$L_{loc} = 1 - GIoU \tag{9}$$

where B_g and B_p indicate ground-truth and predicted bounding boxes, and C_{gp} is the smallest enclosing area regarding two boxes .

To improve imbalance problem in one-stage detector, focal loss [9] provides an effective solution by adaptively reweighting the contribution of easy sample and attending the hard example more. It can be mathematically expressed as (10):

$$\begin{aligned} fl(y_t) = \alpha_0(1 - y_t)^\gamma log(y_t + \epsilon) \\ + (1 - \alpha_0)(y_t)^\gamma log(1 - y_t + \epsilon) \end{aligned} \tag{10}$$

where $fl(y_t)$ denotes standard focal loss function regarding the probability of positive prediction y_t. Moreover, α_0 is the weight factor to balance the number of positive and negative samples (α_0=0.25); γ indicates the modulating term that controls the speed of weight descent of easy samples (γ=2). To ensure the stability of loss, we also add a small constant $\epsilon = 1e - 8$ to avoid loss nan error.

We repurpose focal loss for bounding-box confidence (L_{conf}) and classification loss (L_{cls}) computation in SA-YOLOv3. More specifically, we find high objectness threshold is harmful for detection performance, and thus it would be set to 0.3 in this work. And the weighting factor α is modified by considering the effect of positive prediction, which would be more sensitive to the trend of number of positive and negative samples. Moreover, y_t would be squezed by sigmoid

TABLE II

EVALUATION RESULTS OF ALGORITHMS ON KITTI BENCHMARK

Algorithms	Car			Pedestrian			Cyclist			mAP (%)	FPS	Input size
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard			
SINet [20]	**98.78**	**90.19**	79.24	**88.21**	*79.09*	70.44	**94.34**	**87.12**	*77.30*	**84.97**	25.41	1920 × 576
SSD [24]	87.34	87.74	77.27	50.38	48.41	43.46	48.25	52.31	52.13	60.79	33.17	512 × 512
RFBNet [29]	87.31	87.27	**84.44**	66.16	61.77	58.04	74.89	72.05	71.01	73.66	42.88	512 × 512
ASSD [41]	89.28	*89.95*	*82.11*	69.07	62.49	60.18	75.23	76.16	72.83	75.28	33.04	512 × 512
YOLOv3 [8]	84.37	77.69	75.62	82.58	76.29	73.36	85.14	80.07	77.65	79.19	*45.06*	512 × 512
SA-YOLOv3	88.74	83.67	78.40	84.26	77.53	*75.64*	86.82	80.35	80.79	81.77	**49.22**	512 × 512
SA-YOLOv3	*91.71*	87.16	80.97	*86.49*	**80.63**	**76.91**	*87.28*	*83.44*	**79.83**	*83.82*	32.68	800 × 800

TABLE III

EVALUATION RESULTS OF ALGORITHMS ON BDD100K BENCHMARK

Algorithms	mAP(%)	FPS	Input size
SINet [20]	9.22	20.18	1920 × 576
SSD [24]	14.13	27.64	512 × 512
RFBNet [29]	14.37	42.31	512 × 512
CFENet [30]	**19.10**	22.42	512 × 512
ASSD [41]	15.82	27.09	512 × 512
YOLOv3 [8]	14.63	*44.75*	512 × 512
SA-YOLOv3	*17.26*	**48.47**	512 × 512

function in bbox confidence prediction. Mathematically, α, y_t and L_{conf} can be formulated as (11)-(13): where x_i indicates the prediction vector, and n is the number of samples.

$$\alpha = \alpha_0 \times y_t + (1 - \alpha_0) \times (1 - y_t) \tag{11}$$

$$y_t = \sum_{i=1}^{n} \frac{1}{1 - exp(-x_i)} \tag{12}$$

$$\begin{aligned} L_{conf}(y_t) = \alpha(1 - y_t)^{\gamma} log(y_t + \epsilon) \\ + (1 - \alpha)(y_t)^{\gamma} log(1 - y_t + \epsilon) \end{aligned} \tag{13}$$

For bounding-box classification, we basically follow the standard focal loss (10), but softmax function is adopted to normalize y_t. They can be described as (14)-(15)

$$y_t = \frac{exp(x_i)}{\sum_{i=1}^{n} exp(x_i)} \tag{14}$$

$$\begin{aligned} L_{cls}(y_t) = \alpha_0(1 - y_t)^{\gamma} log(y_t + \epsilon) \\ + (1 - \alpha_0)(y_t)^{\gamma} log(1 - y_t + \epsilon) \end{aligned} \tag{15}$$

Consequently, reconstructed loss function (L_{rec}) is totally formulated as (16), and it would address the class imbalance problem and improve the detection accuracy.

$$L_{rec} = -\sum_{i=1}^{n} (L_{loc} + L_{conf} + L_{cls}) \tag{16}$$

IV. Experiments

To demonstrate the effectiveness of our proposed SA-YOLOv3, we perform comprehensive experiments on KITTI and BDD100K benchmarks. Our implementation environment is based on Tensorflow [45], under Ubuntu18.04 and NVIDIA TITAN RTX GPU.

A. Benchmark and Implementation Detail

KITTI [12] is a commonly used dataset for autonomous-driving research and evaluation. It contains 8 classes with 7481 training and 7518 test images totally, and we mainly focus on three classes in this work, i.e. car, pedestrian and cyclist. Moreover, we utilize 10-fold cross-validation method during model training, which randomly selects one-fold training samples for validation and the remains for model training. Finally, the model is evaluated on test set.

BDD100K [13] is the latest published autonomous-driving benchmark. It provides 10 classes with 100,000 images under different illumination condition and various sizes of objects. The training, validation and test data is split by a ratio of 7:1:2. Also, we use training and validation data during model training, and evaluation model performance on test set.

To make a fair comparison, we implement YOLOv3 and SA-YOLOv3 with the official default setting [46]: the initial learning rate is $1e-3$ with an exponential decay policy; Adam optimizer is used with a momentum of 0.9 and a weight decay of $5e-4$. Additionally, several slight modifications are further introduced in the phase of model training:

1) YOLOv3 and SA-YOLOv3 is firstly trained from scratch with 20k iterations on BDD100K training set, which is argued that better performance and data distribution would be obtained for object detection task [47-48]. And subsequently, pretrained model is continued to fine-tune on KITTI and BDD100K for 15k and 10k iterations, respectively. During model training, warmup [49] and cosine reducing schedule [50] are used to scale the learning rate, avoiding gradient explosion or vanishing problem.

2) Data augmentation is applied during model training to enhance the robustness of algorithm, including random crop, random rotation from 0 to π, horizontal flipping with a probability of 0.5, mixup [51], etc. Furthermore, multi-scale training strategy is adopted, and input size scales from 416×416 to 896×896 by an interval of 32 progressively.

3) Anchor box for KITTI and BDD100K training data is also generated by k-means cluster algorithm [52], as shown in Table I.

In terms of evaluation metric, we use average precision (AP) to calculate detection performance in three difficulty level (easy, moderate and hard) for KITTI benchmark, and compute mean average precision (mAP) for all classes on BDD100K under 0.75 IoU threshold.

TABLE IV
ABLATION STUDIES OF SA-YOLOV3 ON KITTI AND BDD100K BENCHMARKS (INPUT SIZE: 512 × 512)

Dilation convolution		✓			✓
Self-attention module (SAM)			✓		✓
Reconstructed loss function				✓	✓
KITTI (mAP %)	79.19	79.71	80.52	79.92	81.77
BDD100K (mAP %)	14.63	15.17	16.21	15.14	17.26

TABLE V
ABLATION STUDIES OF SSD, ASSD AND SA-SSD ON KITTI AND BDD100K BENCHMARKS

Algorithms	KITTI			BDD100K		
	mAP (%)	FPS	Input size	mAP(%)	FPS	Input size
SSD	60.79	33.17	512 × 512	14.13	27.64	512 × 512
ASSD	75.28	33.04	512 × 512	15.82	27.09	512 × 512
SA-SSD	76.16	33.65	512 × 512	16.47	28.21	512 × 512

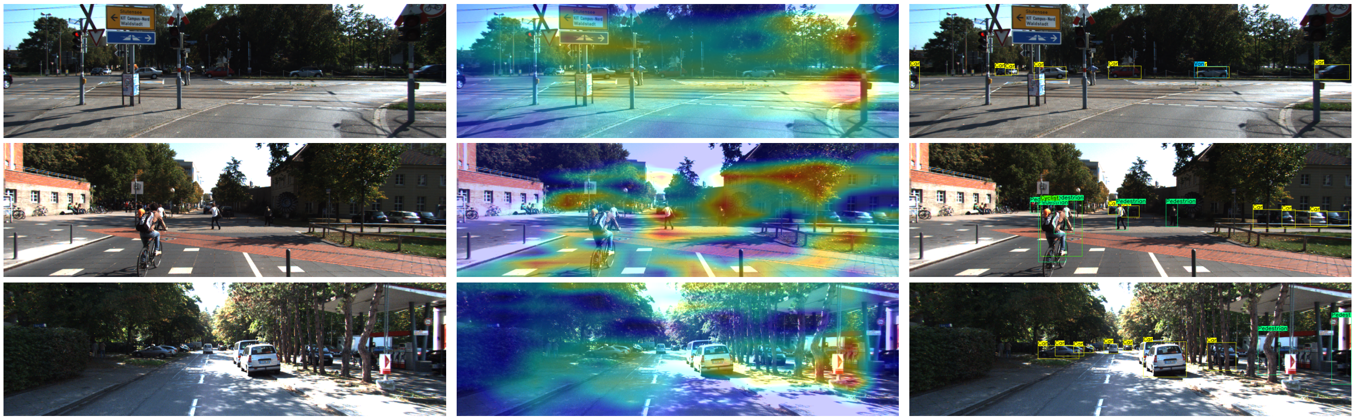


Fig. 5. Visualization results of self-attention feature map on KITTI test set. For each row, the original image, self-attention feature map and detection result are eleborated in left, middle and right column, respectively.

B. Performance Analysis of SA-YOLOv3

We perform experimental analysis of several advanced algorithms on KITTI and BDD100K benchmarks. It is noted that these methods [8, 20, 24, 29, 41] are tested by using pretrained models in their official code; moreover, we don't reproduce the implementation of CFENet [30], and its evaluation result is directly from the original paper.

Table II elaborates detection performance of algorithms on KITTI benchmark. With an input resolution of 512 × 512, SA-YOLOv3 achieves 81.77 mAP at 49.22 FPS, which improves YOLOV3 (79.19 mAP at 45.06 FPS) by 2.58 mAP and 4.16 FPS. In addition, SA-YOLOv3 shows much faster and higher accuracy than other one-stage detectors, e.g. SSD [24], RFBNet [29] and ASSD [41]. When compared with two-stage detector, i.e. SINet [20] that reports 84.97 mAP at 25.41 FPS with an input size of 1920 × 576, SA-YOLOv3 also reports competitive performance with a 83.82 mAP at 32.68 FPS using 800 × 800 input.

Furthermore, Table III illustrates evaluation result of algorithms on BDD100K test set. With an input size of 512 × 512, SA-YOLOv3 has a mAP of 17.26 at 48.47 FPS, which increases YOLOv3 (14.63 mAP at 44.75 FPS) by 2.63 mAP and 3.72 FPS. Also, it outperforms other one-stage methods by a large margin. As for two-stage detector, e.g. CFENet (19.1 mAP at 22.42 FPS) that achieves the 2nd result on BDD100K road object detection challenge, SA-YOLOv3 presents a slightly inaccurate performance but much more efficient, indicating better trade-off in terms of accuracy and inference.

Consequently, experimental results validate the superiority and competitiveness of our proposed SA-YOLOv3, and better trade-off between accuracy and inference speed makes it more suitable for autonomous driving.

C. Ablation Study

We conduct ablation studies to explore the contribution of different components in our proposed algorithm, including dilated convolution, self-attention module (SAM) and reconstructed loss function. As elaborated in TABLE IV, the introduction of SAM significantly improves detection accuracy by 1.33 and 1.58 mAP on KITTI and BDD100K, respectively. Furthermore, dilated convolution and reconstructed loss function also provide considerable contribution for performance improvement, which validates the effectiveness of each component in our proposed SA-YOLOv3.

Additionally, to further validate the flexibility and superiority of SAM, we also introduce SAM between feature map and predicted layer into SSD architecture, and it is termed as SA-SSD. To make a fair comparison, we only use attentive mechanism in ASSD, and perform evaluation experiment

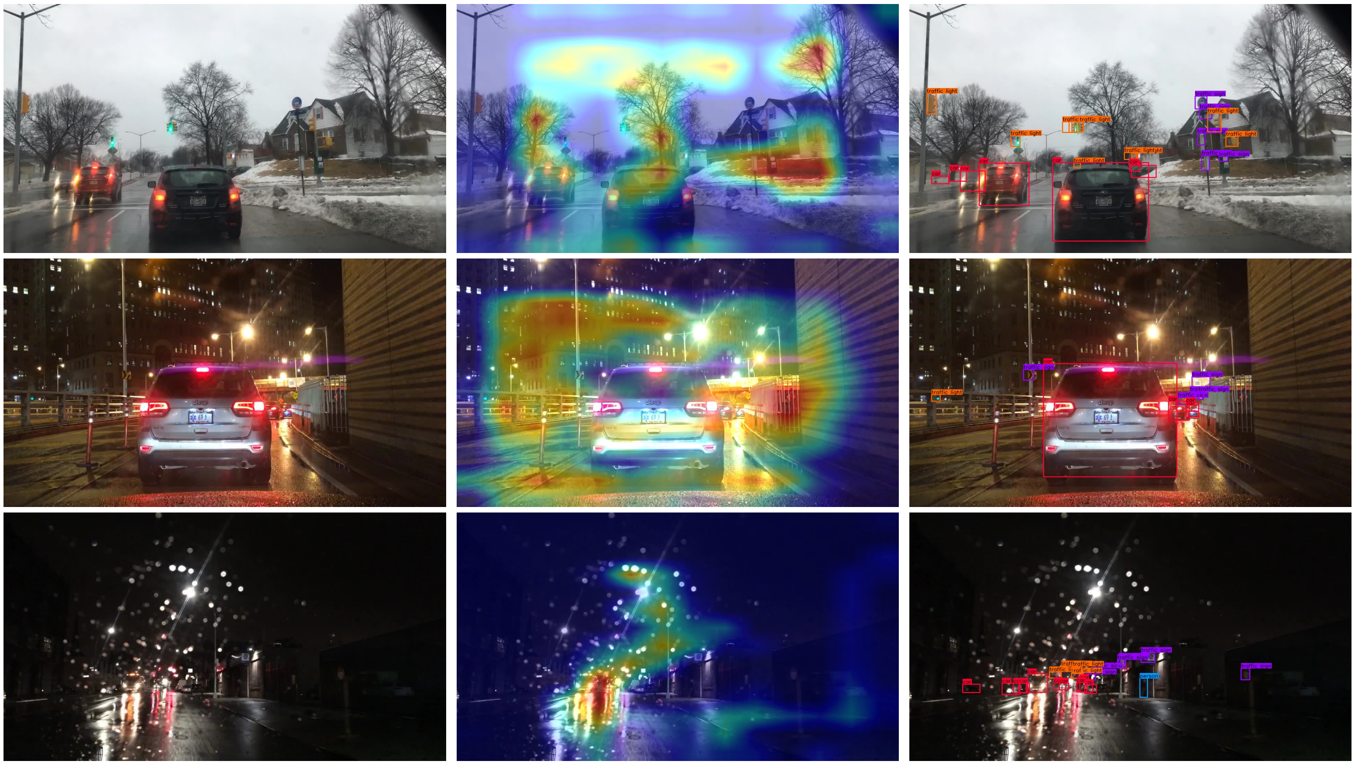


Fig. 6. Visualization results of self-attention feature map on BDD100K test set. For each row, the original image, self-attention feature map and detection result are shown in the first, seccond and third column.

between SSD, ASSD and SA-SSD in the same parameter setting. As shown in Table V, SA-SSD significantly improves SSD by 15.37 mAP on KITTI and 2.34 mAP on BDD100K benchmark. Moreover, SA-SSD also outperforms ASSD by 0.88 and 0.65 mAP on KITTI and BDD100K, respectively. It therefore demonstrates the flexibility and effectiveness of SAM when being plugged into other architecture. More importantly, with the introduction of dilated convolution, SAM can focus on contextual information and learn more discriminative feature, which are also helpful to performance improvement.

D. Visualization Result

We provide a number of visualization results to better understand detection performance of SA-YOLOv3, including self-attention feature map, trend of loss function and object detection result.

1) Self-attention feature map: To investigate how self-attention feature map works in SA-YOLOv3, we follow SAM by an additional softmax probability function, and project heatmap onto the original image to visualize attentive feature. Experiment is conducted on KITTI and BDD100K test set, and visualization result can be observed in Fig.5 and Fig.6. Clearly, self-attention feature map highlights useful region and provides significant feature information for object detection. Taking the first row in Fig.5 for example, with the guidance of attention map, our detector can mainly focus on potential region that may contain targeted objects, and simultaneously suppress negative feature or unrelated background. Also, this benefit helps to alleviate class imbalance problem of one-stage detector to some extent, thus enhancing object detection performance.

2) Trend of loss function: To explore the effect of loss function, we dynamically record the trend of loss function in YOLOv3 and SA-YOLOv3 during model training, on two benchmarks. As illustrated in Fig.7, the trend of loss function of YOLOv3 greatly fluctuates at the early training step, and usually falls into local optimal point. By contrast, SA-YOLOv3 performs more stable training, and converges much faster during model training, which demonstrates the effectiveness of reconstructed loss. Furthermore, better convergence performance indicates it can make effect on class imbalance problem and would provide detection performance improvement.

3) Object detection result: To better understand performance of algorithms, we further visualize detection result of YOLOv3 and SA-YOLOv3 on both KITTI and BDD100K test data, as illustrated in Fig.8 and Fig.9. Obviously, YOLOv3 easily suffers from negative or mis-localization detection, and reports many inaccurate results. Contrarily, SA-YOLOv3 performs much accurate object localization and greatly decreases negative and false detection. Moreover, our proposed detector can find some hard samples that YOLOv3 cannot. For instance, YOLOv3 doesn't detect some pedestrians in the shadow and occluded cars, while SA-YOLOv3 can recognize these hard objects and perform accurate detection (e.g. the first row in Fig.8 and the last row in Fig.9). It further demonstrates the superiority of our proposed SA-YOLOv3.

V. Conclusion

To explore an autonomous-driving suitable object detector, we propose an efficient but accurate architecture termed as

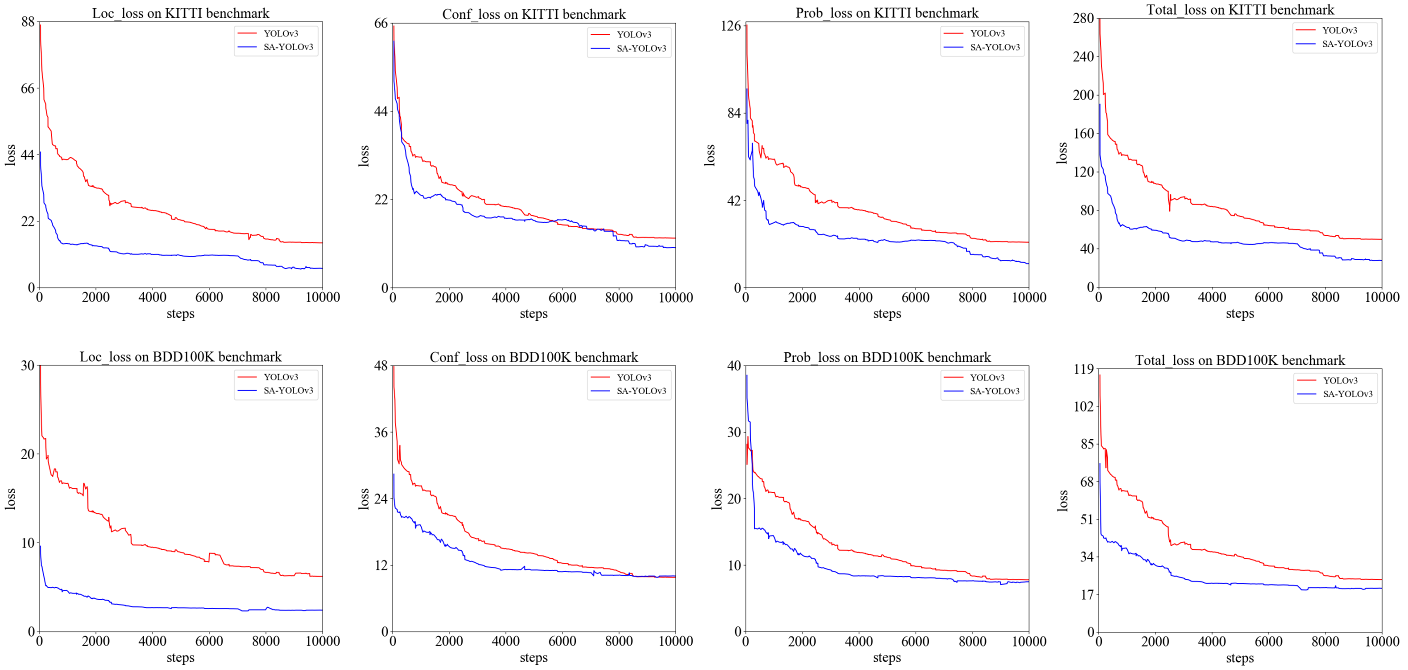


Fig. 7. The trend of loss function of YOLOv3 and SA-YOLOv3 on KITTI and BDD100K benchmarks. It is noted that training loss on KITTI and BDD100K benchmarks are illustrated in the top and bottom row, respectively.

SA-YOLOv3 by incorporating dilated convolution with self-attention module (SAM). Furthermore, loss function based on GIoU and focal loss is also reconstructed for alleviating class imbalance and improving detection performance. A number of validation experiments and ablation studies on KITTI and BDD100K benchmarks are conducted, and evaluation results demonstrate the superiority and competitiveness of SA-YOLOv3: compared with YOLOv3, it can significantly improve detection accuracy by 2.58 mAP on KITTI and 2.63 mAP on BDD100K, with real-time inference (more than 40 FPS). Also, SA-YOLOv3 reports better trade-off in terms of detection accuracy and speed than other state-of-the-art detectors, indicating its suitability for autonomous-driving application. Moreover, ablation studies further present the effectiveness of dilated convolution and self-attention module (SAM) in our proposed algorithm, and visualization results provide a novel perspective to investigate what self-attention feature map presents and how loss function changes. In future work, we would enhance the robustness and practicality of our algorithm in more data from real-world scene. Also, exploration of uncertainty estimation in object detection is a hot issue for autonomous driving.

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China under Grant No. 61672082 and 61822101, Beijing Municipal Natural Science Foundation No. 4181002.

References

[1] B. Wu, A. Wan, F. Iandola, et al., ”SqueezeDet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving,” in *CVPR*, 2017.

[2] X. Dai, ”Hybridnet: A fast vehicle detection system for autonomous driving,” *Signal Processing: Image Communication*, vol. 70, pp. 79-88, 2019.

[3] R. Girshick, J. Donahue, T. Darrel and T. Malik, ”Rich feature hierarchies for accurate object detection and semantic segmentation,” In *CVPR*, 2014.

[4] R. Girshick, ”Fast r-cnn,” In *ICCV*, 2015.

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

[6] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, ”You only look once: Unified, real-time object detection,” In *CVPR*, 2016.

[7] J. Redmon and A. Farhadi, ”Yolo9000: Better, faster, stronger,” In *CVPR*, 2017.

[8] J. Redmon and A. Farhadi, ”Yolov3: An incremental improvement,” *arXiv preprint arXiv: 1804.02767*, 2018.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, et al., ”Focal loss for dense object detection,” In *ICCV*, 2017.

[10] F. Yu, ”Multi-scale context aggregation by dilated convolutions,” In *ICLR*, 2016.

[11] H. Rezatofighi, N. Tsoi, J.-Y. Gwak, A. Sadeghian, I. Reid and S. Savarese, ”Generalized Intersection over Union: A metric and a loss for bounding box regression,” In *CVPR*, 2019.

[12] A. Geiger, P. Lenz and R. Urtasun, ”Are we ready for autonomous driving?” In *CVPR*, 2012.

[13] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, et al., ”Bdd100k: A diverse driving video database with scalable annotation tooling,” *arXiv preprint arXiv: 1805.04687*, 2018.

[14] N. Dalal and B. Triggs, ”Histograms of oriented gradients for human detection,” In *CVPR*, 2005.

[15] Z. Sun, G. Bebis and R. Miller, ”Monocular precrash vehicle detection: Features and classifiers. *IEEE Transactions on Image Processing*, vol. 15, no. 7, pp. 2019-2034, 2006.

[16] Q. Yuan, A. Thangali,V. Ablavsky and S. Sclaroff, ”Learning a family of detectors via multiplicative kernels,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.33, pp. 514-530, 2011.

[17] P. Sermanet, K. Kavukcuoglu, S. Chintala and Y. Lecun, ”Pedestrian detection with unsupervised multi-scale feature learning,” In *CVPR*, 2013.

[18] Y. Li, K. He, J. Sun, et al., ”R-fcn: Object detection via region-based fully convolutional networks,” In *NIPS*, 2016.

[19] T-Y. Lin, P. Dollar, R. Girshick, K, He, et al., ”Feature pyramid networks for object detection,”In *CVPR*, 2017.

[20] X. Hu, X. Xu, Y. Xiao, et aL., ”SINet: A scale-insensitive convolu-

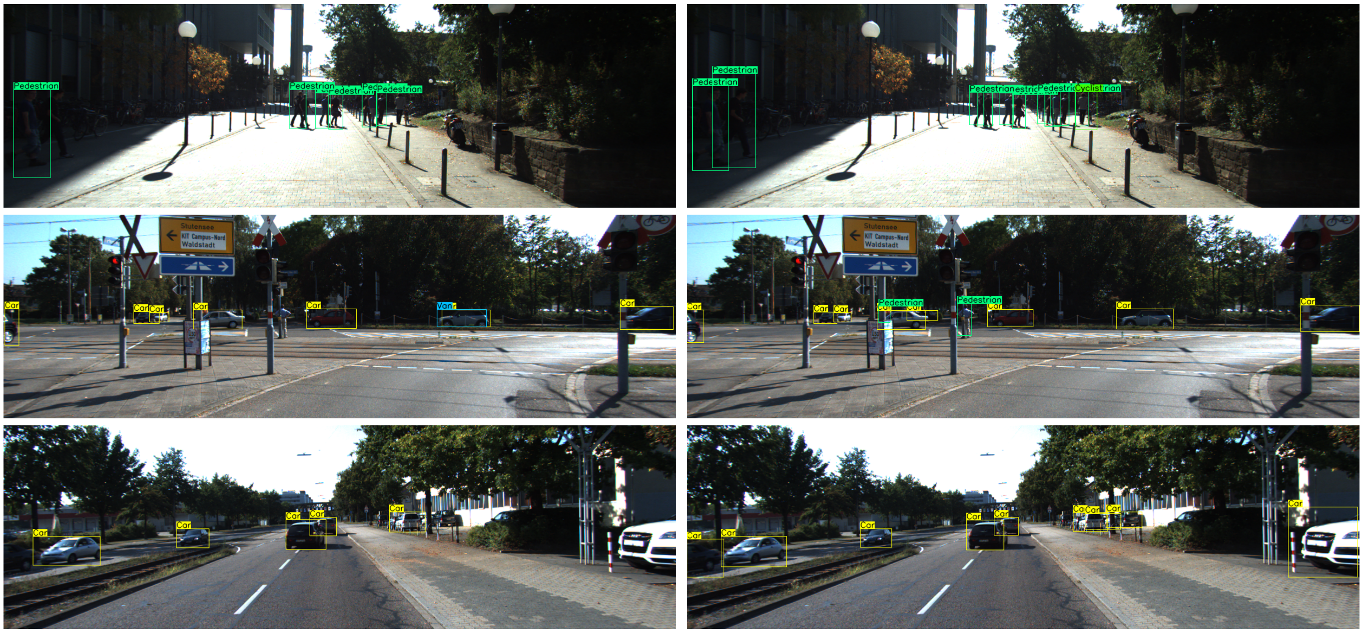


Fig. 8. Detection results of YOLOv3 and SA-YOLOv3 on KITTI test set. The left column shows detection results of YOLOv3, whereas the right column presents detection results of SA-YOLOv3.

tional neural network for fast vehicle detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 1010-1019, 2019.

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

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," In *ICML*, 2015.

[23] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," In *CVPR*, 2016.

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, et al., "Ssd: Single shot multibx detector," In *ECCV*, 2016.

[25] Z.-X. Li and F.-Q. Zhou, "FSSD: Feature fusion single shot multibox detector," In *CVPR*, 2017.

[26] Z.-Q. Shen, Z. Liu, J.-G. Li, Y.-G. Jiang, Y.-R. Chen and X.-Y. Xue, "Dsod: Learning deeply supervised object detectors from scratch," In *CVPR*, 2017.

[27] S.-F. Zhang, L.-Y. Wen, X. Bian, Z. Lei and S.-Z. Li, "Single-shot refinement neural network for object detecton," In *CVPR*, 2018.

[28] P. Zhou, B.-B. Ni, C. Geng, J.-G. Hu and Y. Xu, "Scale-transferrable object detection," In *CVPR*, 2018.

[29] S. Liu, D. Huang and Y. Wang, "Receptive field block net for accurate and fast object detection," In *ECCV*, 2018.

[30] Q. Zhao, T. Sheng, Y. Wang, F. Ni and L. Cai, "Cfenet: An accurate and efficient single-shot object detection for autonomous driving," In *ACCV*, 2018.

[31] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv: 1409: 0473*, 2014.

[32] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," In *NIPS*, 2017.

[33] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova K, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv: 1810.04805*, 2018.

[34] K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho and Y. Bengio, "Attention-based models for speech recognition," In *NIPS*, 2015.

[35] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," In *ICML*, 2015.

[36] J.-S. Lu, J.-W. Yang, D. Batra and D. Parikh, "Hierarchical question-image co-attention for visual question answering," In *NIPS*, 2016.

[37] H. Zhang, I. Goodfellow, D. Metaxas and A. Odena, "Self-attention generative adversarial networks," In *ICML*, 2019.

[38] V. Mnih, N. Heess, A. Graves and K. Kavuk, "Recurrent models of visual attention," In *NIPS*, 2014.

[39] H. Hu, J. Gu, Z. Zhang, J. Dai and Y.Wei, "Relation networks for object detection," In *CVPR*, 2018.

[40] X. Wang, R. Girshick, G. Gupta and K. He, "Non-local neural networks," In *CVPR*, 2018.

[41] J.-R. Yi, P.X. Wu and D. Metaxas, "Assd: Attentive single shot multibox detector," *Computer Vison and Image Understanding*, vol. 189, pp. 102827, 2019.

[42] M. Holschneider, M.-R. Kronland, J. Morlet and P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," *Wavelets*, pp. 286-297, 1990.

[43] M.-J Shensa, "The discrete wavelet transforms: wedding the atrous and mallat algorithm," *IEEE Transactions on Signal Processing*, vol. 44, no. 10, pp. 2464-2482, 1992.

[44] P.-Q. Wang, P.-F. Chen, Y. Yuan, D. Liu, Z.-H. Huang, X.-D. Hou and G. Cottrell, "Understanding convolution for semantic segmentation," In *WACV*, 2017.

[45] M. Abadi, P. Barham, J.-M. Chen, A. Davis, et al., "Tensorflow: A system for large-scale machine learning," *arXiv preprint arXiv: 1605.08695*, 2016.

[46] Redmon, *https://pjreddie.com/darknet/yolo/*.

[47] K.-M. He, R, Girshick and P. Dollar, "Rethinking imageNet pre-training," *arXiv preprint arXiv: 1811.08883*, 2018.

[48] Z.-Q. Shen, Z. Liu, J.-G. Li, Y.-G. Jiang, Y.-R. Chen and X.-Y. Xue, "Object detection from scratch with deep supervision," In *ICLR*, 2018.

[49] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," In *ICLR*, 2017.

[50] Z. Zhang, T. He, H. Zhang, Z.-Y. Zhang, J.-Y. Xie, M. Li, "Bag of freebies for training object detection neural network," *arXiv preprint arXiv: 1902.04103*, 2019.

[51] H. Zhang, M. Cisse, Y.-N. Dauphin and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," In *ICLR*, 2018.

[52] J. Choi, D. Chun, H. Kim and H.-J. Lee, "Gaussian yolov3: An accurate ad fast object detector using localization uncertainty for autonomous driving," In *ICCV*, 2019.

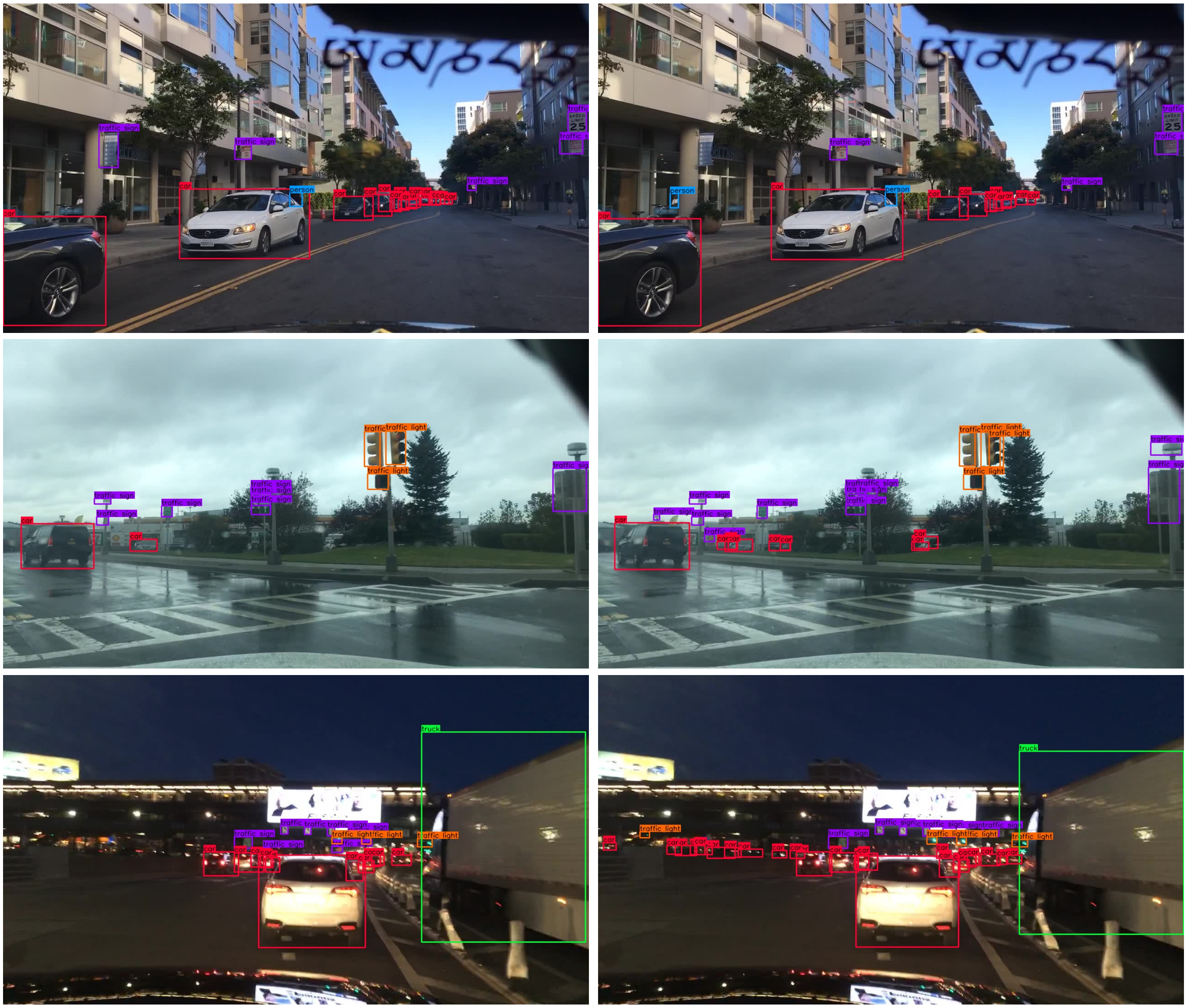


Fig. 9. Detection results of YOLOv3 and SA-YOLOv3 on BDD test set. The first column elaborates detection results of YOLOv3, while the second column demonstrates detection results of SA-YOLOv3.

Daxin Tian [M'13, SM'16] is currently a professor in the School of Transportation Science and Engineering, Beihang University, Beijing, China. He is IEEE Senior Member, IEEE Intelligent Transportation Systems Society Member, and IEEE Vehicular Technology Society Member, etc. His current research interests include mobile computing, intelligent transportation systems, vehicular ad hoc networks, and swarm intelligent.

Chunmian Lin is currently working towards the Ph.D degree with the School of Transportation Science and Engineering, Beihang University, Beijing, China. His current research interests include autonomous driving, image processing, computer vision, artificial intelligence and deep learning, particularly their applications in intelligent transportation systems.

Jianshan Zhou received the B.Sc. and M.Sc. degrees in Traffic Information Engineering and Control from Beihang University in 2013 and 2016, respectively. He is currently working towards the Ph.D. degree with the School of Transportation Science and Engineering, Beihang University, Beijing, China. His current research interests include wireless communication, artificial intelligent system, and intelligent transportation systems.

Xuting Duan received the Ph.D degree in Traffic Information Engineering and Control from Beihang University, Beijing, China. He is currently an assistant professor with the School of Transportation Science and Engineering, Beihang University. His current research interests include vehicular ad hoc networks, cooperative vehicle infrastructure system and internet of vehicles.

Yue Cao received the Ph.D degreee from University of Northumbria at Newcastle. He has been a senior lecturer in University of Northumbria and an international lecturer in Lancaster University, respectively. Currently, he is a professor in the School of Transportation Science and Engineering, Beihang University, Beijing, China. His current research interests include connected and automated vehicles, information security in CVIS, traffic planning and charge management in new energy vehicles.

Dezong Zhao received the B.Eng. and M.Sc. degrees in Control Science and Engineering from Shandong University in 2003 and 2006 respectively, and Ph.D degree in Control Science and Engineering from Tsinghua University in 2010. He is currently an assistant professor in the Department of Aeronautical and Automotive Engineering, Loughborough University, United Kingdom. His current research interests include connected and automated vehicles, autonomous vehicles, machine learning and dynamic optimisation.

Dongpu Cao is an associate professor in the Department of Mechanical and Mechatronics Engineering, University of Waterloo, Canada. He is also the Canada Research Chair in Driver Cognition and Automated Driving, and Director of Waterloo Cognitive Autonomous Driving (CogDrive) Lab. Prof. Cao has received the SAE Arch T. Colwell Merit Award in 2012, and three Best Paper Awards from the ASME and IEEE conferences. Currently, he serves as an Associate Editor for IEEE Transactions on Vehicular Technology, IEEE Transactions on Intelligent Transportation Systems, IEEE/ASME Transactions on Mechatronics, IEEE Transactions on Industrial Electronics, etc. His research interests include automated driving, cognitive autonomous driving, driver cognition, driver-automation collaboration, vehicle dynamics and control.