

## Precise Single-stage Detector

Aisha Chandio <sup>†1</sup>, Gong Gui<sup>†1</sup>, Teerath Kumar<sup>† 2</sup>, Irfan Ullah <sup>3</sup>, Ramin Ranjbarzadeh<sup>4</sup>, Arunabha M Roy<sup>5</sup>, Akhtar Hussain<sup>6</sup>, and Yao Shen <sup>\*1</sup>

<sup>1</sup>*Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China.*

<sup>2</sup>*Department of Software Engineering, School of Computing, National University of computer and emerging sciences, Islamabad 44000, Pakistan*

<sup>3</sup>*School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, 44000, Pakistan*

<sup>4</sup>*School of Computing, Faculty of Engineering and Computing, Dublin City University, Dublin ,Ireland.*

<sup>5</sup>*Aerospace Engineering Department, University of Michigan, Ann Arbor, MI 48109, USA*

<sup>6</sup>*Department of Information Technology, Quaid-e-Awam University of Engineering, Science, and Technology, Nawabshah,Pakistan*

---

<sup>†</sup>These authors contributed equally to this work.

<sup>\*</sup>Corresponding author. Email addresses: yshen@cs.sjtu.edu.cn

## Abstract

*Background and objectives:* Deep learning (DL) logarithms have shown an impressive performance in various tasks. Among them, Single-stage object detectors (SSD) mainly depends on classification network to extract features, multiple feature maps to predict, and classification confidence to guide the filtration of the overlapping prediction boxes. However, there are still two problems causing some inaccurate results: (1) In the process of feature extraction, with the layer-by-layer acquisition of semantic information, local information is gradually lost, resulting into less representative feature maps; (2) During the Non-Maximum Suppression (NMS) algorithm due to inconsistency in classification and regression tasks, the classification confidence and predicted detection position cannot accurately indicate the position of the prediction boxes. *Methods:* In order to address these aforementioned issues, we propose a new architecture, a modified version of Single Shot Multibox Detector (SSD), named Precise Single Stage Detector (PSSD). Firstly, we improve the features by adding extra layers to SSD. Secondly, we construct a simple and effective feature enhancement module to expand the receptive field step by step for each layer and enhance its local and semantic information. Finally, we design a more efficient loss function to predict the IOU between the prediction boxes and ground truth boxes, and the threshold IOU guides classification training and attenuates the scores, which are used by the NMS algorithm. *Main Results:* Benefiting from the above optimization, the proposed model PSSD achieves exciting performance in real-time. Specifically, with the hardware of Titan Xp and the input size of 320 pix, PSSD achieves 33.8 mAP at 45 FPS speed on MS COCO benchmark and 81.28 mAP at 66 FPS speed on Pascal VOC 2007 outperforming state-of-the-art object detection models. Besides, the proposed model performs significantly well with larger input size. Under 512 pix, PSSD can obtain 37.2 mAP with 27 FPS on MS COCO and 82.82 mAP with 40 FPS on Pascal VOC 2007. The experiment results prove that the proposed model has a better trade-off between speed and accuracy.

Keywords: Objection Detection(OD); Precise Single Stage Detector (PSSD); Deep-Convolutional Neural Networks (D-CNNs); Deep Learning (DL); Machine Learning (ML).

## 1. Introduction :

In recent years, deep learning algorithms have became a powerful tool which can automatically capture nonlinear and hierarchical features that has shown great success on various applications, in particular, image domains such as classification, segmentation, detection, captioning, and various others [1, 2, 3, 4, 5, 6]. Furthermore, it has been extended for different classification tasks including audio classification [7, 8, 9, 10], text classification [11, 12, 13, 14, 15, 16], various signals classification [17, 18, 19, 20], and multi-modality object classification [21, 22, 23], event detection [24, 25], and various other applications [26, 27, 28]. Among them, object detection [29, 30, 31, 32, 33, 34] has been a central interest to vast majority of researchers. To this end, there are various algorithms such as YOLO, fast-RCNN, faster-RCNN, and others that have been successfully used for object detection over the years. Deep learning algorithms for object detection has gained significant attention over the last few decades [1, 2, 3]. Object detection, aiming at locating object instances from a large number of predefined categories in natural images, is one of the most basic and challenging problems in computer vision [35, 36]. With the rapid development of CNN, object detection has made remarkable progress and gradually turns into two main architectures: two-stage and single-stage. Two-stage algorithms such as Faster Recurrent Neural Network (FRCNN) [37, 38] where the first stage only distinguishes a large number of background regions and obtains rough object proposals without considering the specific class of the object. It has then followed by the second stage that classifies each proposal and optimizes the location according to the features extracted from CNN network [39, 40]. Due to the existence of refinement conducted by the second stage, two-stage algorithms cannot achieve performance in real-time [40, 38]. Therefore, single-stage algorithms have been the major priority for various object detection applications due to real-time detection [35, 36] and thus, it is the particular interest of the current work. The single-stage algorithms directly perform classification and location optimization based on the default boxes [40, 38]. For example, You Look Only Once (YOLO) [41] and SSD [42] have achieved fast real-time detection speed but simultaneously sacrificed detection accuracy. In recent years, single-stage detectors are improving their accuracy, but still, they cannot achieve a better trade-off between speed

and accuracy [38].

In this paper, on the premise of guaranteeing the real-time performance of the model, we propose a novel architecture named Precise Single Stage Detector (PSSD) based on the original SSD that provides solutions to the two key questions: (1) How to enrich the information of features used by predictors without relying on a deep backbone of the model like ResNet-101 [43]? (2) Whether it is reasonable to rely on classification confidence to determine the filtration of overlapping boxes in the process of the NMS algorithm? In the next subsections, we will address the aforementioned questions in order to establish the main challenges and corresponding improvements necessary that we have implemented in our proposed PPSD model.

### **1.1 Feature richness:**

Considering the significant overhead caused by the image pyramid, SSD [42] puts forward a feature pyramid to solve the problem of multi-scale detection. The deep features in the classification network contain more semantic information, which is suitable for identifying large objects, while the shallow low-level features are more suitable for small objects. However, the lack of semantic information in shallow features and the loss of local details in deep features deteriorate the precision of SSD. Feature Pyramid Networks (FPN) [44] adds deep semantic information to shallow features using a U-shaped structure to obtain a more effective feature pyramid, which improves the effect of small object detection. DetNet [45] combines dilated convolution to reduce local information loss by decreasing the down-sampling steps so that the positioning accuracy of large objects is improved. From the above-mentioned networks, it is noted that the features used in each scale predictor need not only suitable semantic information but also local texture information for more accurate positioning. The information richness of each level feature is significant to the detection effect. But the problem is how to construct a high-performance feature pyramid with as little overhead as possible.

### **1.2 Filtration of overlapping boxes:**

Usually, to prevent the results from overlapping, we set the NMS as the final operation

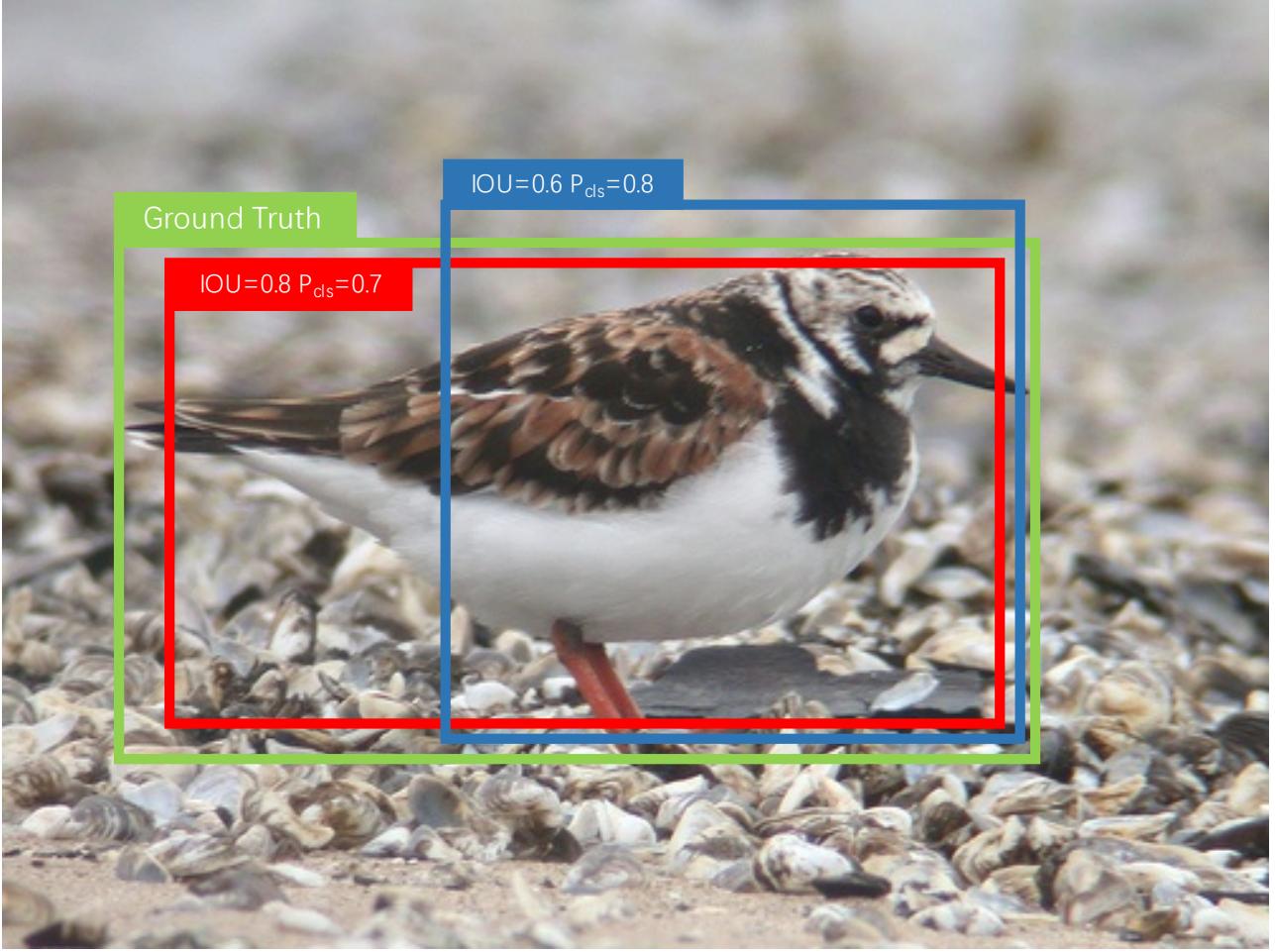


Figure 1: A example of test results without NMS, the blue box will be retained because of higher classification confidence.

of object detection. In the NMS algorithm, the prediction boxes with the highest classification confidence are retained and other boxes are filtered when IOU between the two is greater than the threshold. As shown in Figure 1, this may lead to inaccurate results. IOUNet [46] guides the NMS to alleviate this problem by predicting the IOU between the regression boxes and their ground truth. Herein, the key question arises - how to make it more effective in single-stage algorithms?

### 1.3 Contribution:

To improve the detection effect, we mitigate the above problems based on the SSD model. Firstly, we improve features by introducing extra layers in the SSD to make the basic feature pyramid

more suitable for feature fusion. Secondly, we send these features to the feature enhancement module (FEM). The FEM consists of two sub-modules, the receptive field expansion module (RFM) and two-way FPN. The RFM is used for expanding the receptive field of each feature step-by-step and two-way FPN is employed to supplement more local and semantic information. Finally, a new loss function is designed to predict the IOU between the regression boxes and their ground truth. Then we use the IOU label to guide the training of classification to improve the consistency of classification and regression tasks and use prediction IOU to guide inference to attenuate the scores of low-quality boxes.

We have validated our method on MS COCO [47] and Pascal VOC [48] benchmarks. Under the hardware of Titan Xp with the input size of 320 pix, the proposed PSSD can obtain 33.8 mAP with 45 FPS on COCO and 81.28 mAP with 66 FPS on Pascal VOC 2007 outperforming state-of-the-art object detection models. Besides, our model also performs significantly well with a larger input size. Under 512 pix, PSSD can obtain 37.2 mAP with 27 FPS on MS COCO and 82.82 mAP with 40 FPS on Pascal VOC 2007. The experiment results prove that our model has a better trade-off between speed and accuracy. The rest of this paper is organized as follows.: Section 2 covers related literature relevant to this work; the proposed methodology has been described in section 3; the new loss function introduced in this study has been detailed in section 4; section 5 contains experimental results; Section 6 deals with the relevant finding and discussion of the proposed model; the conclusions and prospects of the current work have been discussed in section 7.

## **2. Related Work :**

### **2.1 Object detection architectures:**

With the introduction of R-CNN [49], the deep learning method of object detection begin to develop which illustrated superior performance in object detection compared to traditional methods. To this end, various approaches such as DetNet [45], Cascade R-CNN [50] have been proposed. For each image, the R-CNN uses selective search [51] to create roughly 2000 region proposals. To provide more accurate candidate boxes of arbitrary sizes rapidly and limit the

searching space in object recognition, the selective search method relies on simple bottom-up (BU) grouping and saliency cues [52, 53]. Each region proposal suggestion is twisted or cropped to a fixed resolution and the CNN network [54] is used to extract the high-level, robust, and semantic 4096-dimensional feature as the final representation. Different region suggestions are assessed on a set of positive areas and negative regions using pre-trained category-specific linear SVMs for various classes. To obtain final bounding boxes for conserved item positions, the scored areas are modified using bounding box regression and filtered using greedy non-maximum suppression (NMS). As R-CNN obtained the fixed and same final representation for each region suggestion by cropping and warping, the item may exist in part in the cropped zone, and the warping procedure may result in undesirable geometric distortion. The accuracy of recognition will be harmed by these content losses or distortions, especially when object scales differ.

To mitigate this problem, SPP-NET [55] suggested a new CNN architecture called SPP-net based on the principle of spatial pyramid matching (SPM) [56]. SPM divides the image into a variety of divisions using several finer to coarser scales and accumulates these quantized local features into mid-level representations. Unlike R-CNN, SPP-net projects region suggestions of variable sizes to fixed-length feature vectors using feature maps from the fifth convolutional layer. The feature maps are reused because they include not only the intensity of local responses but also correlations with their spatial placements [55]. The SPP layer is the next layer following the final convolutional layer. Taking a three-level pyramid, if the number of feature maps in the convolutional layer is 256, the final feature vector for each suggested region acquired after the SPP layer has a size of  $256 \times (12 + 22 + 42) = 5376$ . Even though SPP-net outperforms R-CNN in terms of efficiency and accuracy, it still has several significant flaws. Feature extraction, network fine-tuning, SVM training, and bounding-box regressor fitting are all part of SPP-multistage net’s workflow, which is nearly identical to R-CNN’s. As a result, extra storage space is still necessary. Furthermore, the fine-tuning technique proposed in [55] cannot update the convolutional layers before the SPP layer. As a result, it’s expected that the accuracy of very deep networks drops. Consequently, Girshick [49] suggested a new CNN architecture called Fast R-CNN, which included a multitask loss on classification and bounding box regression. The entire image is processed with convolutional layers to yield feature maps, similar to SPP-Net. Then, from each suggested region with an ROI, a fixed-length feature

vector is extracted. The RoI pooling layer is a subset of the SPP layer, which consists of only one pyramid level. Following that, each feature vector is fed into a series of FC layers before branching off into two sibling output layers. One output generates softmax probabilities for all the classes including the background while the second output layer uses four real-valued numbers to encode revised bounding-box coordinates. All parameters in these operations are optimized end-to-end using a multitask loss (except for the development of region suggestions). In order to build candidate boxes using biased sampling [57], the latest object identification networks use various approaches to generate a set candidate pool of isolated region suggestions, such as selective search and Edgebox. The computation of region suggestion is also a hurdle in increasing efficiency. [37] proposed an extra region proposal network (RPN) that functions in a practically cost-free manner by sharing full-image convolutional features with the detection network to overcome this problem. RPN is accomplished using an FCN, which predicts both object bounds and scores simultaneously at every single position. Faster R-CNN [37] utilizes a network named Region Proposal Network(RPN) to generate region suggestions. Using anchors, this module learns to determine if an object is present at its associated position for each feature map. RPNs have a distinct and trainable module that predicts a set of bounding boxes considering reference boxes of various aspect ratios and sizes. As a result, this model defines a collection of anchors with varying aspect ratios and sizes on the relevant input images for each position on the feature extractor output. Although, on one hand, Faster R-CNN allows region proposal-based CNN architectures for object detection to be trained in an end-to-end manner, on the other hand, its training is highly time-intensive, and RPN creates objectlike regions together with the background rather than object instances, and it is not capable of coping with objects of extreme dimensions or forms. The R-FCN’s last convolutional layer, Unlike Faster R-CNN, creates a total of  $k^2$  position-sensitive score maps with a fixed grid of  $k \times k$  for each class. These are then appended to the position-sensitive RoI pooling layer’s output in order to get accumulative output. Finally, softmax responses across each class are generated by averaging  $k^2$  position-sensitive scores in each RoI to form a  $C + 1 - d$  vector. To achieve class-agnostic bounding boxes, another  $4k^2 - d$  convolutional layer is added. More powerful classification networks can be used with RFCN to achieve object detection in a fully convolutional architecture by sharing nearly all of the layers. In order to improve scale

invariance, several object identification systems have used feature pyramids generated from the image [53, 55] at the cost of an increase in both training time and memory usage. Feature pyramid can extract rich semantics from all layers and be trained end to end with all scales. In a deep ConvNet, the in-network feature hierarchy provides feature maps with various spatial resolutions while also introducing huge semantic gaps due to varying depths.

To obtain a real-time detector, Single-stage architecture is investigated. SSD [42] and YOLO series [41] are the most representative. However, their accuracies are lower than Two-stage algorithms. In recent years, Single-stage detectors have been improving, RetinaNet [58], which uses more powerful classification networks, predictors, and dense anchors and proposes focal loss to alleviate the problem of several easy samples in the training process and enables Single-stage architecture accuracy to be aligned with Two-stage accuracy. In addition, in terms of lightweight Single-stage architecture, RFB-Net [59], RefineDet [60], etc., improve detection accuracy by expanding the receptive field or learning more accurate default boxes.

## 2.2 Feature fusion:

To solve the problem of multi-scale object detection, SSD [42] proposes a feature pyramid. Then the feature fusion becomes popular. FPN [44] supplements the semantic information from top to bottom to improve the effect of small object detection. PA-Net [61] builds another bottom-up feature pyramid based on FPN and adds horizontal connections to expand shallow and deep information for each layer. Kong et al [62]. propose a new efficient feature pyramid that integrates layer features in a highly nonlinear and efficient manner. RSSD [63], RFBNet [59], DSSD [64], etc. have also conduct feature fusion with different methods. In this paper, we propose a new architecture, a modified version of SSD, named PSSD by introducing extra layers. We use an efficient feature fusion structure to expand the receptive field of each layer and supplement its information, resulting in less redundancy and better results.

## 2.3 IOU prediction:

The drawback of using the classification confidence to filter the overlapping regression boxes is

that a more accurate regression box is filtered out when its classification confidence is not the highest in the NMS algorithm. Based on the Two-stage architecture, IOUNet [46] achieves exciting results by predicting the IOU confidence of the regression boxes and using them to guide the regression and NMS. We optimize the same issues in the single-stage architecture by designing a simple IOU-guided prediction structure.

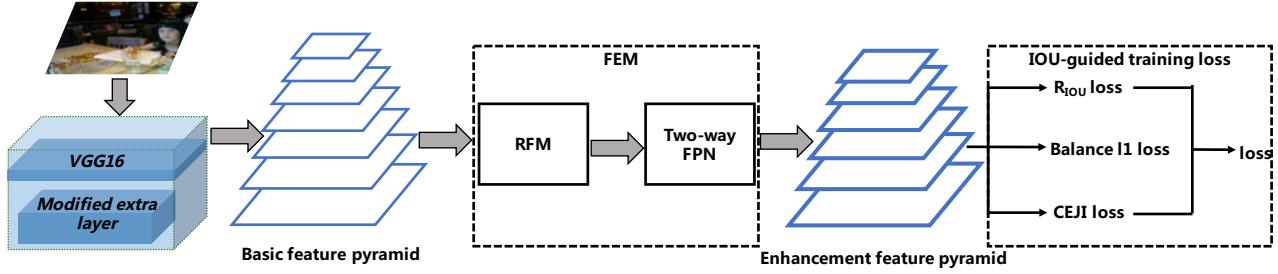


Figure 2: The overall network of the proposed PSSD with SSD backbone optimization, the introduction of a receptive field expansion module (RFM), implementation of two-way FPN and IOU-guided training loss to improve and optimize accuracy and detection speed.

### 3. Methodology

In the present study, our proposed PSSD model has been modified based on the original SSD model in order to improve and optimize the accuracy and detection speed. The overall proposed PSSD architecture has been shown in Figure 2 and each modification has been elaborated in the following subsections.

#### 3.1 SSD backbone optimization

The backbone of SSD consists of VGG16 [65] and extra layers as shown in Figure 3. First of all, we reproduce the SSD algorithm and improve it with the following modifications: removing the L2 normalization of Conv<sub>4-3</sub> layer, adding BatchNorm [66] for each convolution layer, and replacing smooth  $l_1$  with the balance  $l_1$  [67] loss function. To improve the effect of small object detection, we have tried a four-layer FPN structure. However, the experimental results showed

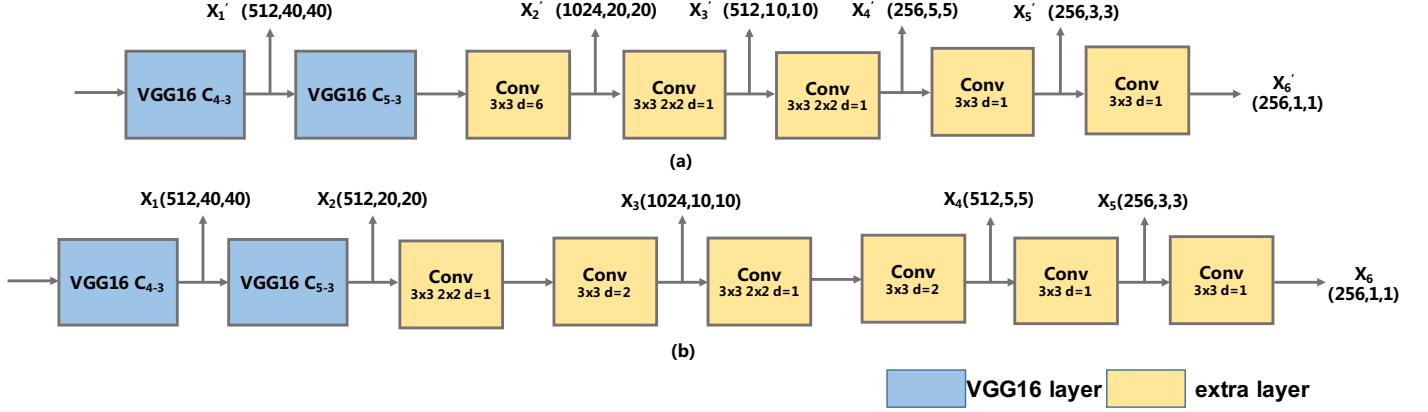


Figure 3: Enhancement of backbone layers (showing convolutional layers only) : (a) original extra layers of SSD; (b) our modified extra layers in PSSD for optimization of uniform basic feature maps distribution in the receptive field which is suitable for multi-scale feature fusion structures.

that the improvement is not as significant as expected. Thus, we have defined the feature maps which are obtained from the SSD backbone and used by predictors as the basic feature maps. The part of the SSD backbone is shown in Figure 3 (a). The basic feature maps are  $X'_1$ ,  $X'_2$ ,  $X'_3$ ,  $X'_4$ ,  $X'_5$ ,  $X'_6$ . To get  $X'_2$ ,  $X'_1$  performs three  $3 \times 3$  convolutions and one of  $3 \times 3$  dilated convolution with dilation of 6. The large dilation size leads to the loss of some of the object semantic and local information from  $X_1$  to  $X_2$ . Meanwhile, the receptive field expansion of basic feature maps used by FPN is uneven. The expansion ratio of  $X'_1$  to  $X'_2$  is much larger than that of  $X'_2$  to  $X'_3$  and  $X'_3$  to  $X'_4$ . Therefore, the effect of FPN is not apparent as expected. So, we redesign the extra layers under the following conditions.

- The receptive field of basic feature maps should expand evenly.
- The parameters of extra layers without the pre-trained model should not increase drastically.
- The coverage of the receptive field at each feature map should be appropriately expanded.

The new extra layers introduced in the proposed model have been shown in Figure 3 (b). We use the  $3 \times 3$  convolution with dilation of 2 instead of two  $3 \times 3$  convolutions. In this way, the extension ratio of the receptive field from  $X_2$  to  $X_3$  and  $X_3$  to  $X_4$  are similar to  $X_1$  to  $X_2$  with a reduced number of network parameters. Considering the third aspect, we use the FEM module (see section 3.2 for details) to expand the receptive field of each feature map without

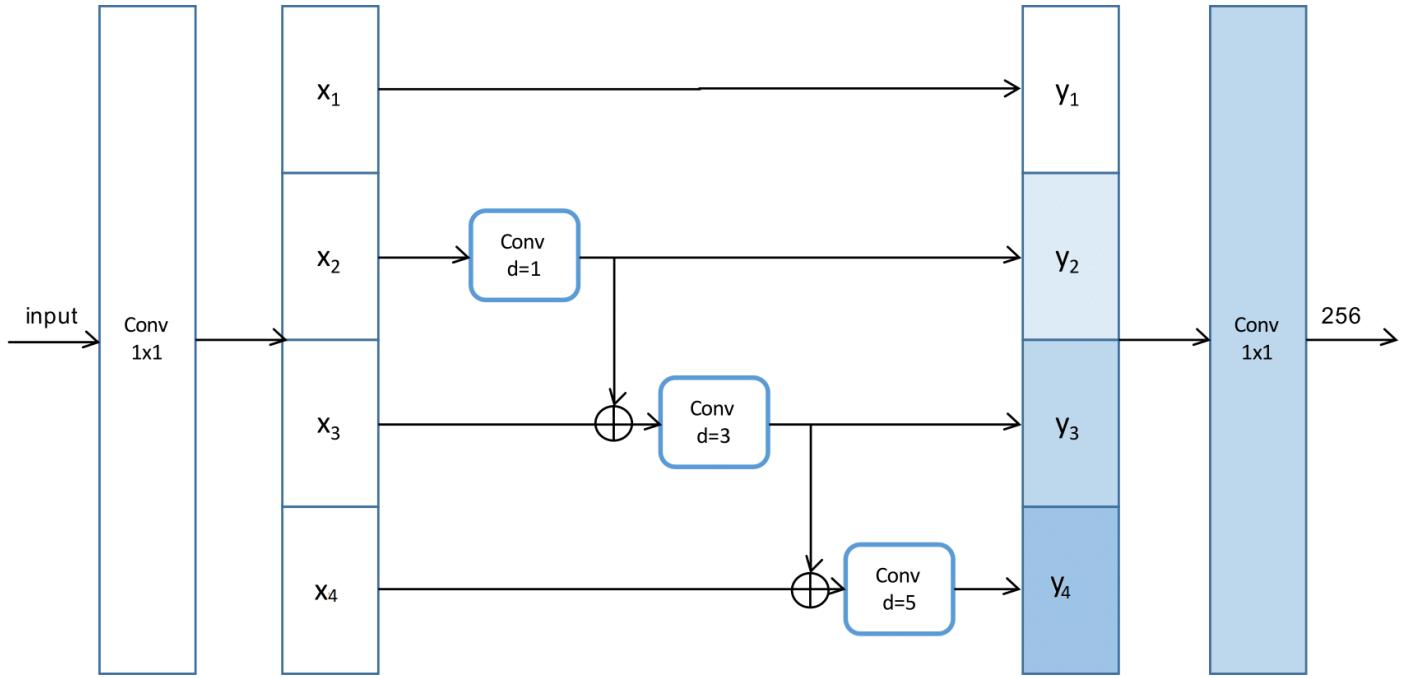


Figure 4: The proposed network structure of receptive field expansion module (RFM) for efficient feature enhancement in PSSD.

affecting others. Overall, with this optimization, the receptive field distribution of basic feature maps is now more uniform and suitable for multi-scale feature fusion structures such as FPN.

### 3.2 Feature Enhancement Module

#### 3.2.1 Receptive field expansion module:

The success of RFB-Net [59] indicates that it is effective to use dilated convolution to expand the receptive field. Considering the effectiveness of the greater receptive field, we design a simple but efficient receptive field expansion module (RFM). As shown in Figure 4, to prevent large dilation size from influencing the calculation speed, we adopt the separation residual structure. First of all, we use  $1 \times 1$  convolution to perform feature conversion on the original feature map  $X$ , and then divide the  $X$  into four blocks:  $X_1, X_2, X_3$ , and  $X_4$ . Where  $X_1$  maps directly to the final feature  $Y_1$  to prevent information loss;  $Y_2$  is gotten by operating a  $3 \times 3$ ,  $d=1$  convolution on  $X_3$ , then add it with  $X_3$  and send the sum to  $3 \times 3$ ,  $d=3$  convolution get  $Y_3$ . Similarly, we conduct a  $3 \times 3$ ,  $d=5$  convolution on the sum of  $Y_3$  and  $X_4$  to get  $Y_4$ . Finally, we

use  $1 \times 1$  convolution to perform cross-channel information integration and dimension reduction on the extended feature  $Y$ . The separated residual design ensures that the receptive field of basic features is largely expanded in the case where the number of parameters does not increase significantly and the original feature details are kept as much as possible. This module is not expensive, except for the last layer, the input and output dimensions of each layer are the same [68].

### 3.2.2 Two-way FPN:

The design of the feature pyramid determines the feature richness and detection performance. FPN introduces high-level semantic information into a shallow layer to enrich features. However, the information on high-level features is also insufficient because of the loss of local details during feature extraction. Unlike DetNet [45], RSSD [63], PANet [61], etc., which bring higher dimension or more convolution computation, we have implemented a simple two-way FPN to alleviate this problem. As shown in Figure 5, after RFM, we construct two information flows. Firstly, we introduce traditional FPN from the  $Y_4$  feature to construct downward semantic information flows and get  $S_1, S_2, S_3$ , and  $S_4$ . Secondly, we construct upward local information flow from the shallower layer of VGG C<sub>3</sub> – 3 to get  $L_1, L_2, L_3, L_4, L_5, L_6$ . Finally, we combine the  $S$  features with the  $L$  features and use a  $3 \times 3$  convolution layer for feature conversion. To reduce overhead and retain the information of each feature, we use 256 dimensions to construct information flow, bilinear interpolation to do up-sampling, and Avg-pooling to do down-sampling. Finally, 512 dimension features are generated for training with stronger predictors.

## 4. Loss Settings based on IOU

### 4.1 IOU branch:

The classification and regression tasks are relatively independent, classification confidence cannot accurately represent the predicted box positions. Hence, some more accurate prediction

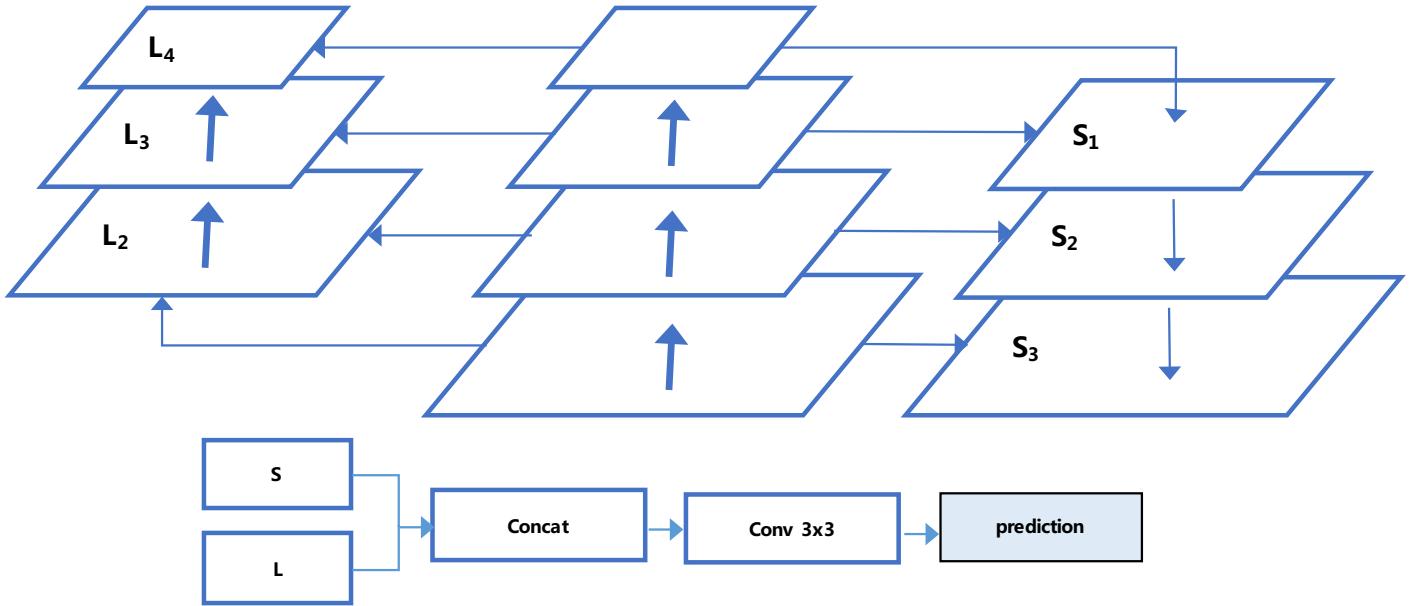


Figure 5: The network structure of Two-way FPN in the proposed PSSD for enhancement of feature richness and reduction of convolution computation for optimized detection performance.

boxes are filtered by NMS. We use a less expensive category-independent way to do IOU-based attenuation for scores that are used in NMS. As shown in Figure 6, we add a branch to predict the IOU between the regression boxes and their ground truth. The results are gotten by directly applying  $3 \times 3$  convolution on the enhanced feature maps, and then using the sigmoid function to normalize it to 0 - 1. We calculate the IOU between the regression boxes and the ground truth. To ensure the convexity of loss and increase the gradient value, we design a new IOU

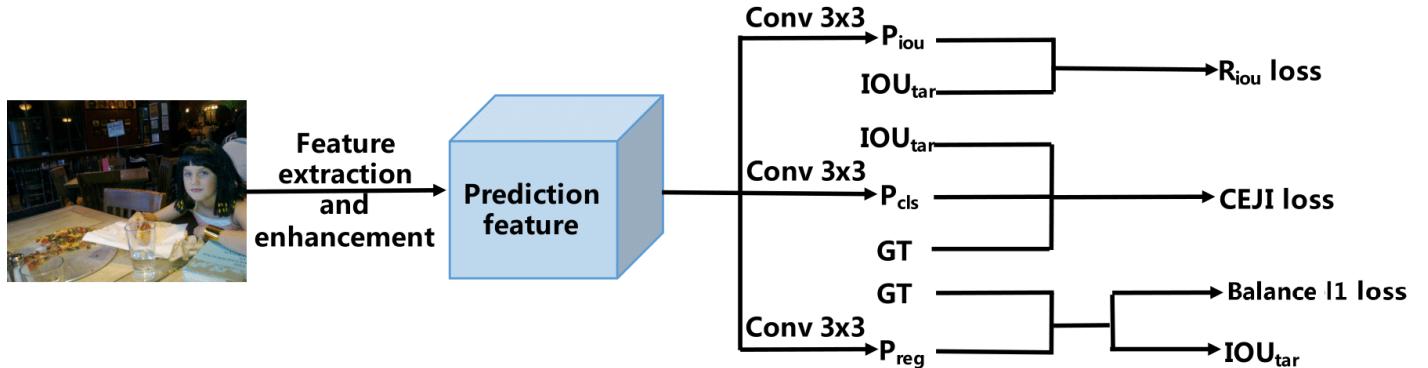


Figure 6: Proposed IOU-guided prediction structure in PSSD that consist of IOU regression loss ( $R_{iou}$  loss), cross-entropy of joint IOU ( $CEJI$  loss), and balance  $l_1$  loss.

regression loss ( $R_{IOUloss}$ ) as follows:

$$R_{IOUloss} = \begin{cases} -\ln \frac{P_{IOU}}{IOU_{tar}} & P_{IOU} < IOU_{tar} \\ -\ln \frac{IOU_{tar}}{P_{IOU}} & P_{IOU} \geq IOU_{tar} \end{cases}$$

where  $P_{IOU}$  is predicted intersection over union and  $IOU_{tar}$  is target intersection over union are precision of IOU and target IOU, respectively.

#### 4.2 Regression Loss:

In order to make large objects regress better, we use balance  $l_1$  ( $L_b(x)$ ) which proposed by Libra-RCNN [67]:

$$L_b(x) = \begin{cases} \frac{a}{b}(b|x| + 1)\ln(b|x| + 1) - \alpha|x| & |x| < 1 \\ \gamma|x| + C & otherwise \end{cases}$$

$$\gamma = \alpha \ln(b + 1)$$

where the default parameters are set as  $\alpha = 0.5$  and  $\gamma = 1.5$ .

#### 4.3 Cross-entropy of joint IOU loss:

For general Single-stage detection algorithms, such as RetinaNet [58], the samples whose IOU is in the range of 0.4-0.5 are ignored, because it is difficult for the model to learn classification boundaries. However, the following problems still exist:

- There are some false positive samples (the IOU of regression results is less than 0.5), and the model still forces them to be identified as objects, which results in low-quality prediction boxes.
- There are some samples whose IOU are in 0.4 - 0.5 that can be learned, ignoring these samples is not conducive to improving the learning ability of the model.
- The traditional cross-entropy loss does not take into account the generation of scores used in NMS during inference.

Thus, we adopt new positive sample rule and design a new cross-entropy of joint IOU ( $CEJI(p, y_{IOUtar})$ ) loss as follows:

$$CEJI(p, y_{IOUtar}) = \begin{cases} -\ln(P_{cls} \cdot IOU_{tar}) & IOU_{tar} \geq 0.5 \\ -\ln(P_{cls}) & IOU_{tar} < 0.4 \end{cases}$$

Where  $IOU_{tar}$  is a gradient value,  $P_{cls}$  is the predicted class probability. we set default boxes whose IOU with their ground truth of more than 0.4 as positive samples. If the default boxes can be regressed more accurately ( $IOU_{tar} \geq 0.5$ ), we calculate the classification loss, otherwise ignore them. Therefore, the model does not force the classification of fake positive samples and can excavate more potential positive samples. Meanwhile, the final optimization goal is changed to maximize  $P_{IOU} \cdot P_{cls}$ , which is consistent with the inference ( $scores = P_{IOU} \cdot P_{cls}$ ). Additionally, considering that balance  $l_1$  is not optimized from the overall of the regression boxes, this loss function can add a monitoring signal about the IOU to regressors to make the prediction results more accurate.

## 5. Experiments

We conduct experiments on the MS COCO [47] and Pascal VOC [48] datasets. We use Cudnn10, two RTX TITANs to train the models, and a Titan Xp to test the results. The pre-trained parameters of VGG16 and VGG16<sub>bn</sub> are public. Our experiments include the following sections:

- Introducing the details of the experiments;
- Demonstrating comparisons of PSSD with other advanced state-of-the-art methods;
- Extensive experiment of PSSD model and demonstrating the validity of each sub-module.

### 5.1 Performance evaluation on Microsoft COCO:

At first, in order to evaluate the performance of PSSD, we perform experiments on the MS COCO dataset that contains 80 categories, 117266 training images, 5k verification images, and 20k test images. This experiment is conducted on the MS COCO *2017train* set. We set the batch size to 32, the size ratio of default boxes to 0.06, 0.15, 0.33, 0.51, 0.69, 0.87, 1.05, the

weight decay to 0.0005 and the momentum of SGD to 0.9. During training, the initial learning rate is set to 0.002, which is reduced by 10 times subsequently at both the 90th and 120th epochs. The total number of training epochs is prescribed as 150. Other settings preserve the original SSD strategy. We conduct the test on MS COCO *2017test-dev* set whose labels are not public and use the coco server to evaluate results. The detection accuracy comparison

Table 1: Performance comparison including detection accuracy, detection speed (in FPS) for various input sizes between state-of-the-art models and proposed PSSD considering different backbones evaluated on MS COCO *2017test-dev* set.

Method	Backbone	Input size	FPS	Avg. Precision, IOU:			Avg. Precision, Area:		
				0.5:0.95	0.5	0.75	S	M	L
Faster R-CNN [37]	VGG16	~1000×600	7.0	24.2	45.3	23.5	7.7	26.4	37.1
Faster R-CNN w FPN [44]	ResNet-101	~1000×600	6.0	36.2	59.1	39.0	18.2	39.0	48.2
R-FCN [69]	ResNet-101	~1000×600	9	29.9	51.9	-	10.8	32.8	45.0
CoupleNet [70]	ResNet-101	~1000×600	8.2	34.4	54.8	37.2	13.4	38.1	50.8
Deformable R-FCN [71]	Inc-Res-v2	~1000×600	-	37.5	58.0	40.8	19.4	40.1	52.5
SSD300 [42]	VGG16	300×300	43	25.1	43.1	25.8	6.6	25.9	41.4
DSSD321 [64]	ResNet-101	321×321	9.5	28.0	46.1	29.2	7.4	28.1	47.6
YOLO <sub>v3</sub> 320 [72]	DarkNet53	320×320	45	28.2	-	-	51.5	-	-
RefineDet320 [60]	VGG16	320×320	38.7	29.4	49.2	31.3	10.0	32.0	44.4
RFB-Net300 [59]	VGG16	300×300	66.7	30.3	49.3	31.8	11.8	31.9	45.9
<b>PSSD320</b>	VGG16	320×320	<b>45</b>	<b>33.8</b>	<b>52.2</b>	<b>35.8</b>	<b>14.8</b>	<b>38.5</b>	<b>50.3</b>
SSD512 [42]	VGG16	512×512	22	28.8	48.5	30.3	10.9	31.8	43.5
DSSD513 [64]	ResNet-101	513×513	53.3	35.2	13.0	35.4	51.1		
YOLO <sub>v3</sub> 608 [72]	DarkNet53	608×608	19.8	33.0	57.9	34.4	18.3	35.4	41.9
RefineDet512 [60]	VGG16	512×512	22.3	33.0	54.5	35.5	16.3	36.3	44.3
RFB-Net512 [59]	VGG16	512×512	33.3	33.8	54.2	35.9	16.2	37.1	47.4
<b>PSSD512</b>	VGG16	512×512	<b>27</b>	<b>37.2</b>	<b>55.9</b>	<b>40.3</b>	<b>18.7</b>	<b>41.6</b>	<b>51.4</b>

between the present PSSD model and other advanced state-of-the-art detectors such as Faster R-CNN [37], Faster R-CNN w FPN [44], R-FCN [69], CoupleNet [70], Deformable R-FCN [71], SSD300 [42], DSSD321 [64], YOLO<sub>v3</sub>320 [72], RefineDet320 [60], RFB-Net300 [59], SSD512 [42], DSSD513 [64], YOLO<sub>v3</sub>608 [72], RefineDet512 [60], RFB-Net512 [59] has been considered for various backbones including VGG16, ResNet-101, Inc-Res-v2, and DarkNet53 as shown in Table 1.

From the comparison, it is worth noting that the PSSD320 achieves 33.8 mAP, surpassing most object detectors with more powerful backbones and larger input sizes. For example, R-FCN performs 29.9 mAP at the input size of 1000×600 pix, and DSSD513 performs 33.2 mAP. Even if

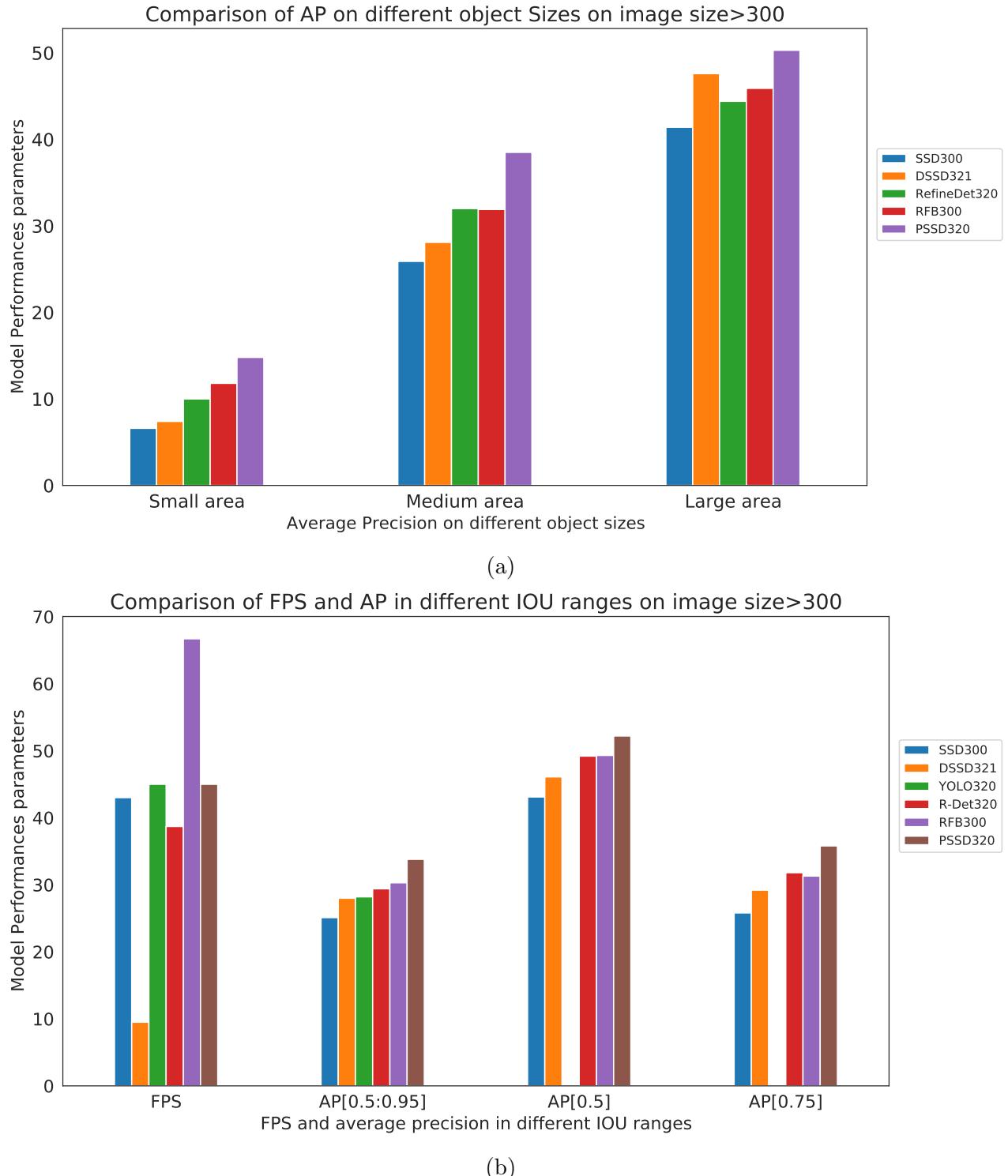


Figure 7: Performance comparison in term of FPS and AP, for different IOU and area using input size more than 300 and less than 321

they use a deeper ResNet-101 [43], the mAP is lower than the PSSD320. At the same time, our model retains a powerful inference speed at 45 FPS based on the optimization. Compared with other real-time methods, PSSD has higher precision. For example, RFB-Net300 achieves 30.3 mAP, and RefineDet320 achieves 29.4 mAP. In addition, under the input size of 512 pix, our model also shows a good performance with 37.2 mAP at 27 FPS. Its accuracy and speed exceed YOLO<sub>v3</sub>608 [72] and sophisticated RetinaNet500 [58]. For better clarity in the comparison, we visualize the result in different groups i.e. average precision in different areas such as small, medium, and large as shown in Fig. 7-(a), it is noticeable that PSSD310 in all area sizes outperformed the performance of the other models. We not only visualize the results from an area perspective but also visualized the results from FPS, and different AP thresholds. As shown in Fig. 7-(b), PSSD312 has consistently outperformed the previous methods using different AP threshold ranges while PSSD312 has the second-highest FPS among all methods. Like 7, we do visualization to check the effect of different input sizes, we change input size from 320 to 512. As it is shown in figure 8 (a), PSSD512 with different area objects has outperformed all other methods, while in terms of FPS, performance is dropped but still comparable, and in terms of different AP threshold ranges, it outperformed all other methods except AP [0.5], where Yolo has demonstrated highest performance as shown in Fig. 8-(b). In summary, our proposed PSSD has shown massive performance gain in terms of different accuracy parameters and detection speed.

Table 2: Average Precision Comparison of Modified SSD.

Method	AP	$AP_{50}$	$AP_{75}$	$AP_{small}$	$AP_{medium}$	$AP_{large}$
SSD&	28.5	46.9	29.9	0.98	30.8	46.2
Modified SSD&	28.8	46.3	30.1	10.9	31.0	45.6
SSD&_FPN	29.7	49.0	30.7	12.2	32.4	45.4
Modified SSD&_FPN	30.5	50.0	31.9	12.7	33.8	46.1

## 5.2 Ablation Study:

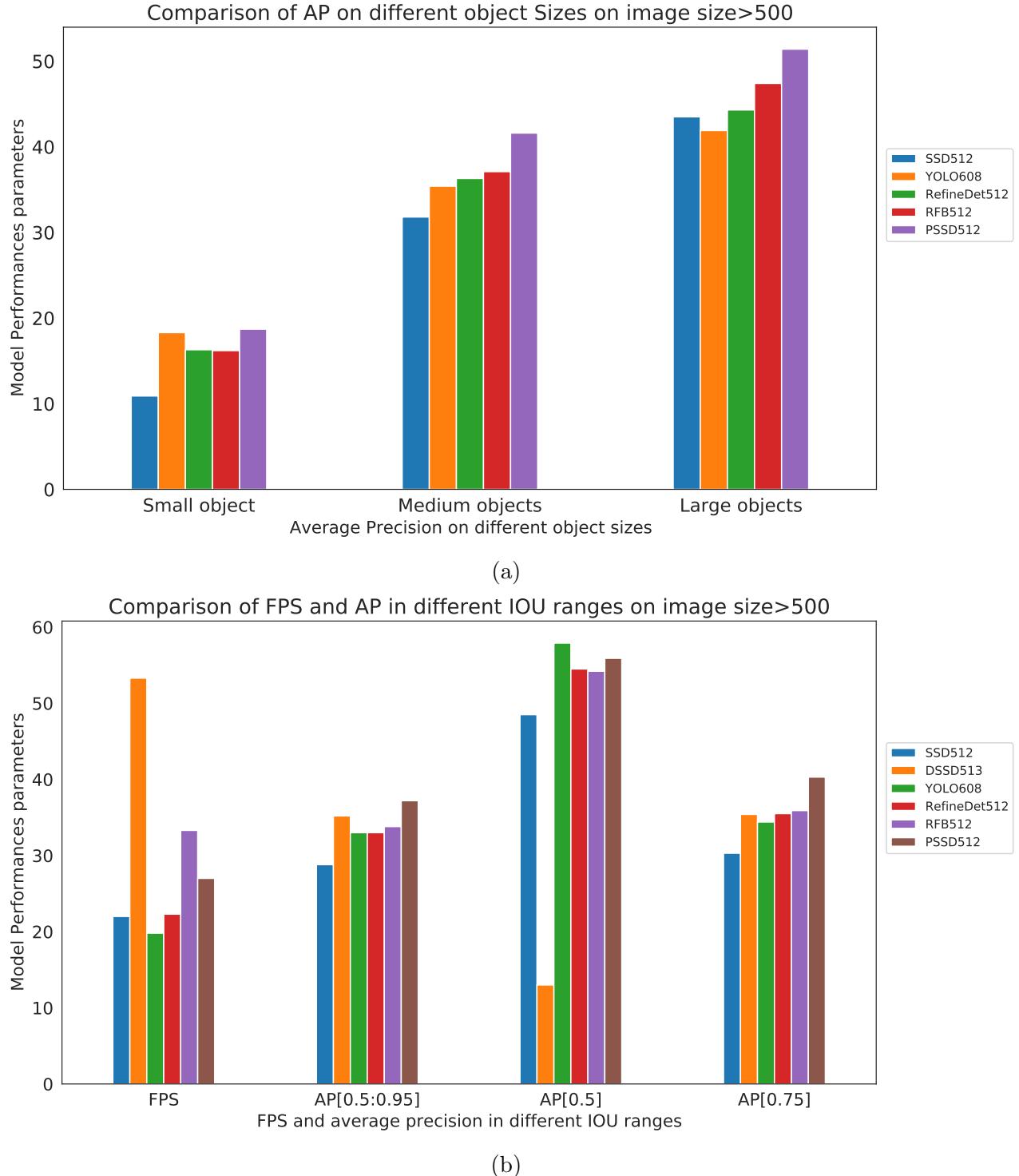


Figure 8: Performance comparison in term of FPS and AP, for different IOU and area using input size more than 500

Table 3: Ablation study on PSSD320: The detection results are evaluated on *2017val* set

+ Bn-Balance <sub>l1</sub>	✓	✓	✓	✓	
+ Two-way FPN		✓	✓	✓	
+ RFM			✓	✓	
+ IOU-guided prediction				✓	
AP	25.8	28.8	31.5	32.2	33.8
$AP_{50}$	44.7	46.3	51.0	51.2	52.2
$AP_{75}$	26.8	30.1	32.9	33.7	35.8
$AP_{small}$	7.2	10.9	13.5	13.8	14.8
$AP_{medium}$	27.4	31.0	34.7	35.7	38.5
$AP_{large}$	41.4	45.6	47.6	48.3	50.3

Since our model consists of a modified SSD backbone and multiple sub-modules, we verify the validity of each module extensively. The baseline is a simple detector of the original SSD [42] with a  $320 \times 320$  input size.

### 5.2.1 Modified Single-shot Multibox Detector SSD<sup>&</sup> backbone:

To verify the effectiveness of the SSD<sup>&</sup> backbone for the feature fusion module, we conduct four experiments with FPN. The results are shown in Table 2. First of all, it makes the original SSD mAP increase by 2.7% that adding BatchNorm and Balance l<sub>1</sub> [67]. The accuracy of SSD<sup>&</sup> after modifying the extra layers is 28.8% which is similar to the original SSD (removing the effect of increasing parameters). However, adding the FPN structure to the modified SSD<sup>&</sup> has a better effect that the mAP reaches 30.5%. The original SSD-FPN is only 29.7%. In contrast, under the FPN structure, the modified backbone is increased by 0.5%. This shows new extra layers are more suitable for the feature fusion structure.

### 5.2.2 FEM module:

The FEM module operation process is divided into two steps. Firstly, the receptive field of basic features is expanded by RFM step by step and then we use Two-way FPN to supplement the more in-depth and shallower information. It can be seen from Table 3 and Table 2 that the Two-way FPN is more effective than the original FPN, especially under the large object

Table 4: Comparisons of different loss function settings.

loss	AP	$AP_{50}$	$AP_{75}$	$AP_{small}$	$AP_{medium}$	$AP_{large}$
$L_2 + CEJI$	33.3	51.9	35.5	14.9	37.8	50.2
$R_{iou} + CE$	33.0	51.7	34.7	15.1	36.5	49.5
$R_{iou} + CEJI$	33.8	52.2	35.8	14.8	38.5	50.3

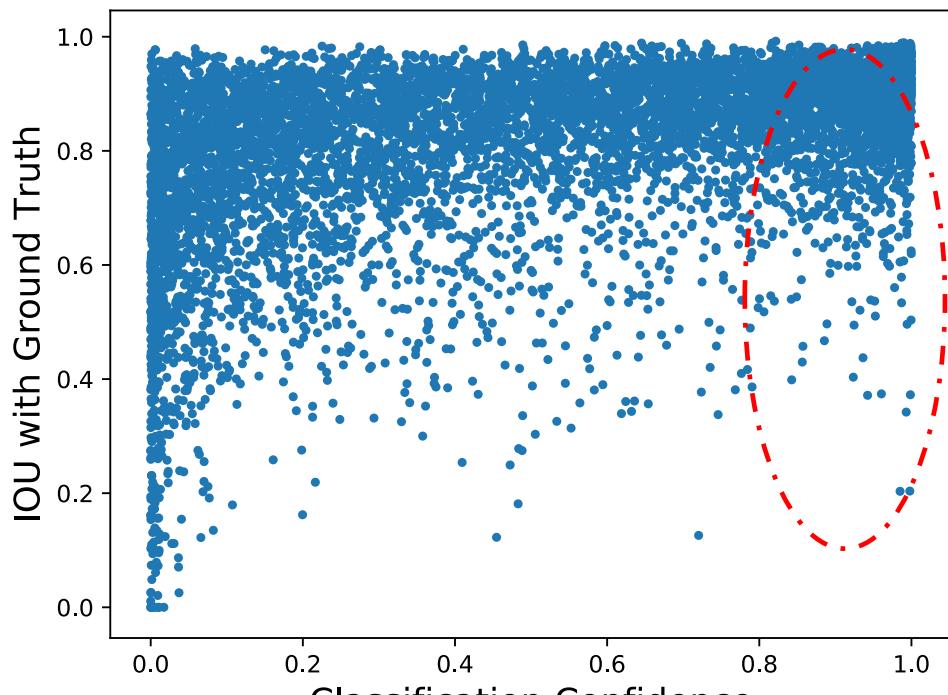
and the  $IOU_{0.75}$  index, the overall reach 31.5% which is 1% higher than the original FPN. The addition of RFM further improves the model’s mAP to 32.2% in line with our expectations. In short, benefiting from the FEM module, the feature information of each feature map is more abundant, and the detection effects of both large and small objects are greatly improved.

### 5.2.3 IOU Prediction structure:

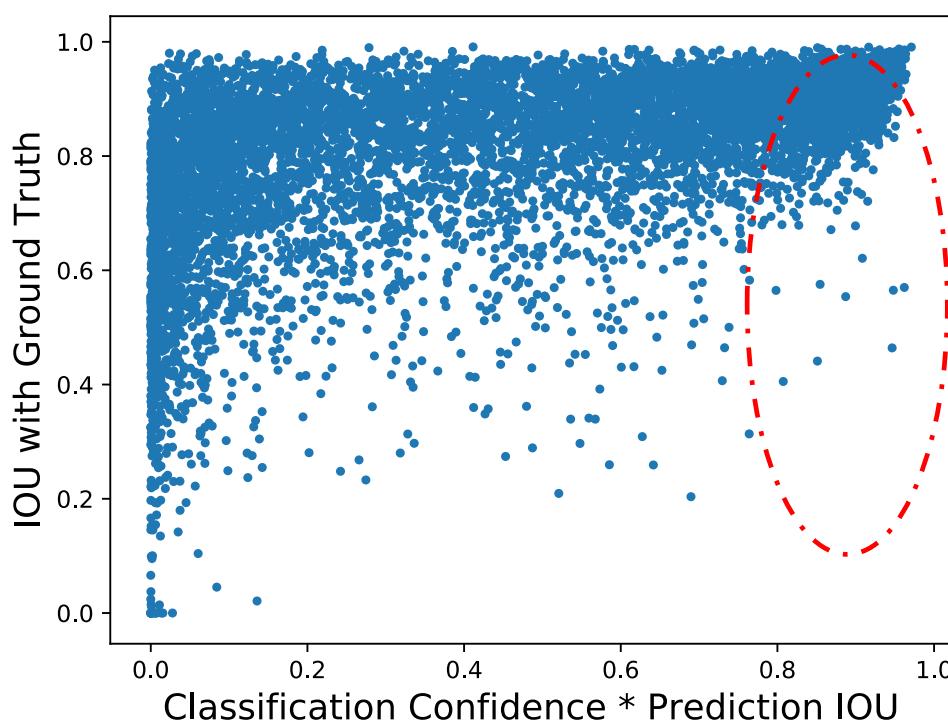
We predict the IOU between the regression boxes and their ground truth with very little overhead and use them in classification loss and inference. Overall, the IOU-guided prediction structure obtains effective results and improves mAP by 1.6% as shown in Table 3. In addition, we verify the validity of  $R_{IOU}$  loss and CEJI loss. As shown in Table 4, Comparing  $R_{IOU}$  loss with  $l_2$  loss, the new loss has a better effect with an increase of 0.5%. At the same time, CEJI loss has increased by 0.8% compared to CE loss. It means that the two new loss functions can both better optimize the model. As can be seen from Fig. 9, the number of prediction boxes that have a high score but low  $IOU_{tar}$  is significantly less, so it is more reasonable to use the new score to conduct overlapping boxes filtration.

## 5.3 Performance evaluation on Pascal VOC 2007

To further validate PSSD, we conduct experiments on a relatively small Pascal VOC dataset. Specifically, we train our model on the union of 2007 *trainval* set and 2012 *trainval* set. This training set contains 20 categories and 16555 images and then we test the result on VOC 2007*test* set which has 5k images. We set the batch size to 32, using the default boxes of 0.06, 0.15, 0.33, 0.51, 0.69, 0.87, 1.05 ratio, and the weight decay of 0.0005, the momentum of 0.9. The initial learning rate is set to 0.002, attenuated by 10 times at 160th and 200th



(a)



(b)

Figure 9: Distribution of NMS scores and real IOUs on MS COCO 2017val set test results. The prediction boxes in the red circle is significantly reduced.

Table 5: Detection accuracy comparisons on Pascal VOC 2007test set.

Method	Backbone	FPS	mAP
Faster R-CNN	VGG16	7	73.2
Faster R-CNN w FPN	ResNet-101	5	76.4
R-FCN	ResNet-101	9	80.5
Deformable R-FCN	ResNet-101	9	82.6
SSD300	VGG16	120	77.2
DSSD321	ResNet-101	9.5	78.6
RefineDet320	VGG16	83	80.5
RFB-Net300	VGG16	40.3	80.0
<b>PSSD320</b>	VGG16	<b>66</b>	<b>81.28</b>
SSD512	VGG16	50	79.8
DSSD513	ResNet-101	5.5	81.5
RefineDet512	VGG16	83	80.5
RFB-Net512	VGG16	24.1	81.8
<b>PSSD512</b>	VGG16	<b>40</b>	<b>82.82</b>

epochs. Finally, the training stops at the 240th epoch. Table 5 shows the comparison of our experimental results with other state-of-the-art (SOTA) models. The performance of PSSD320 is better than that of RefineDet320 [60], YOLO<sub>v3</sub>320 [72] and RFB-Net300 [59]. Its mAP is 81.28% which exceeds the original SSD300 [42] by 4% while maintaining the 66FPS real-time speed. The PSSD512 achieves 82.82% mAP at the speed of 40 FPS, which is superior to most Single-stage and Two-stage object detection systems. It can be seen that our model also shows a good performance on small-scale datasets.

## 6. Discussion

### 6.1 Trade-off between speed and accuracy :

Our model has a better trade-off between speed and accuracy. PSSD has better accuracy and speed than RetinaNet500 [58] and yolo<sub>v3</sub>608 [72]. Meanwhile, PSSD has higher accuracy than RFB-Net [59] and RefineDet [60], it is due to our efficient method optimization. On the one hand, the FEM module makes each feature map used by predictors get richer semantics and local information to achieve better prediction results. On the other hand, in the NMS algorithm, according to to multiply the classification confidence by the predicted IOU as the final scores,

the boxes with high classification confidence but inaccurate regression are more attenuated to keep more accurate prediction boxes as much as possible. The two aspects complement each other to make our model obtain more accurate prediction boxes. At the same time, the proposal of  $R_{iou}$  loss and CEJI loss enable the model to learn better. Considering the inference speed, the design of each module is as lightweight as possible.

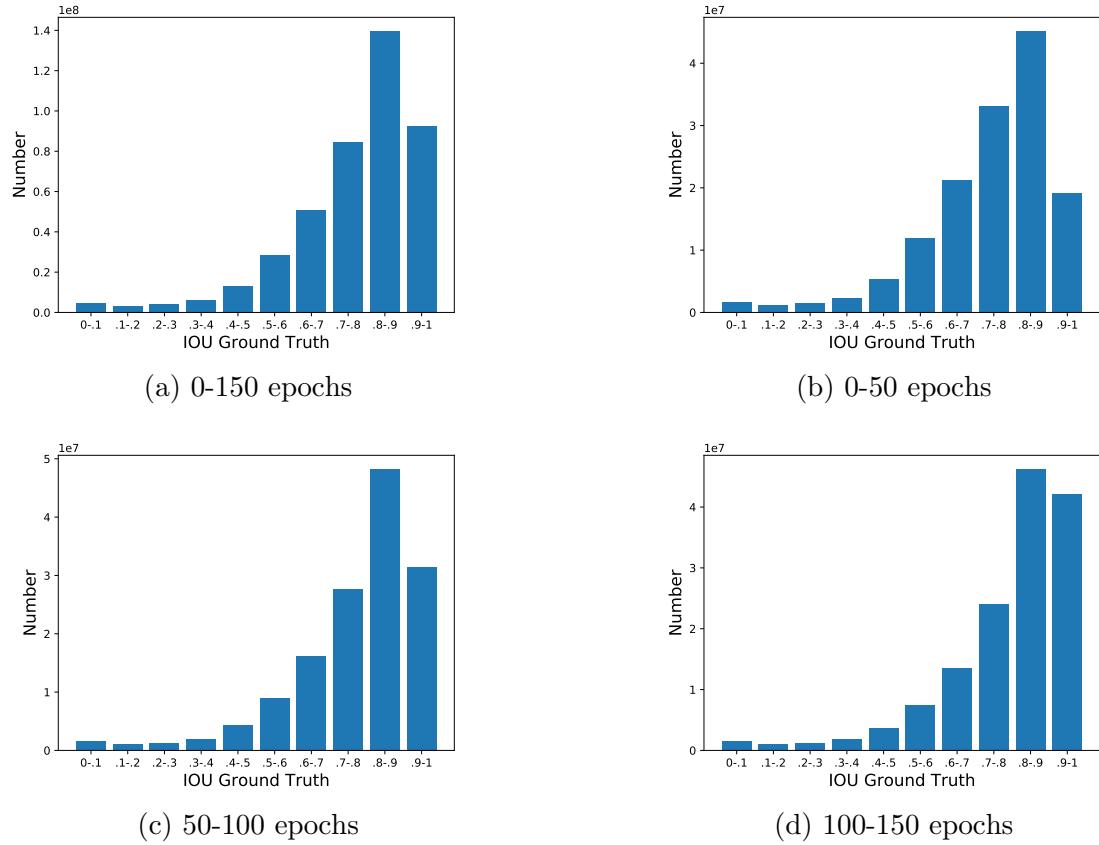


Figure 10: Distribution of  $IOU_{tar}$  at different training stages: (1) Overall unevenness; (2) the IOU samples above 0.9 is significantly increasing as the training deepens.

## 6.2 Existing problems :

A significant problem with Single-stage architecture is the class imbalance between positive and negative samples. Similarly, as shown in Figure 10, this problem is more severe in IOU prediction. There are two situations during the training. Firstly, the distribution of IOU

samples at each interval is uneven. Secondly, the distribution of IOU samples in the training process is constantly changing. These two points lead to a severely imbalanced IOU sample. Therefore, the effect of IOU prediction is unstable, the overall value is high and the accuracy is low. However, in this case, it can still improve by 1.6%, which indicates the prospect of the method, and we have a plan to optimize this problem more in future study.

## 7. Conclusion

In this paper, we address two issues for inaccurate prediction boxes of the Single-stage architecture and proposed new architecture named PSSD and optimize it based on the Single-stage SSD model. Firstly, we introduce the extra layers in the SSD backbone to make it more suitable for high representative feature fusion. Secondly, we use the FEM module to make information on each prediction feature richer and more beneficial to the classification and regression task of each scale. Finally, we design the IOU-guided prediction structure to make the model optimize better. Our proposed architecture showed impressive performance on two popular and large datasets i.e. MS COCO and Pascal VOC and outperformed the previous SOTA methods. Its accuracy even surpasses some of the complex top detectors while maintaining an exciting speed of inference.

**Acknowledgement** This work was partially supported by the Key Program of the National Natural Science Foundation of China under Grant No. 61932014 and the Research and Development Projects of Applied Technology of Inner Mongolia Autonomous Region, China under Grant No. 201802005.

**Conflict of interest** The authors declare that they have no conflict of interest.

**Data availability** The data that support the findings of this study are available to public. Cocodataset link : <https://cocodataset.org/>; PASCAL VOC dataset link: <http://host.robots.ox.ac.uk/pascal/VOC/>

## References

- [1] K. U. Sharma, N. V. Thakur, A review and an approach for object detection in images, International Journal of Computational Vision and Robotics 7 (1/2) (2017) 196--237.
- [2] F. Sultana, A. Sufian, P. Dutta, A review of object detection models based on convolutional neural network, Intelligent Computing: Image Processing Based Applications (2020) 1--16.
- [3] Z. Zou, Z. Shi, Y. Guo, J. Ye, Object detection in 20 years: A survey, arXiv preprint arXiv:1905.05055 (2019).
- [4] T. Kumar, J. Park, M. S. Ali, A. Uddin, S.-H. Bae, Class specific autoencoders enhance sample diversity, Journal of Broadcast Engineering 26 (7) (2021) 844--854.
- [5] T. Kumar, J. Park, M. S. Ali, A. S. Uddin, J. H. Ko, S.-H. Bae, Binary-classifiers-enabled filters for semi-supervised learning, IEEE Access 9 (2021) 167663--167673.
- [6] S. Jamil, M. S. Abbas, A. M. Roy, Distinguishing malicious drones using vision transformer, AI 3 (2) (2022) 260--273.
- [7] J. Park, T. Kumar, S.-H. Bae, Search for optimal data augmentation policy for environmental sound classification with deep neural networks, Journal of Broadcast Engineering 25 (6) (2020) 854--860.
- [8] T. Kumar, J. Park, S.-H. Bae, Intra-class random erasing (icre) augmentation for audio classification, in: Proceedings Of The Korean Society Of Broadcast Engineers Conference, The Korean Institute of Broadcast and Media Engineers, 2020, pp. 244--247.
- [9] A. Chandio, Y. Shen, M. Bendechache, I. Inayat, T. Kumar, Audd: audio urdu digits dataset for automatic audio urdu digit recognition, Applied Sciences 11 (19) (2021) 8842.
- [10] M. Turab, T. Kumar, M. Bendechache, T. Saber, Investigating multi-feature selection and ensembling for audio classification, arXiv preprint arXiv:2206.07511 (2022).
- [11] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, Deep learning-based text classification: a comprehensive review, ACM Computing Surveys (CSUR) 54 (3) (2021) 1-40.

- [12] S. Selva Birunda, R. Kanniga Devi, A review on word embedding techniques for text classification, *Innovative Data Communication Technologies and Application* (2021) 267--281.
- [13] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, D. Brown, Text classification algorithms: A survey, *Information* 10 (4) (2019) 150.
- [14] M. Razno, Machine learning text classification model with nlp approach, in: COLINS 2019. Volume II: Workshop, 2019.
- [15] N. Aslam, I. Ullah Khan, F. S. Alotaibi, L. A. Aldaej, A. K. Aldubaikil, Fake detect: A deep learning ensemble model for fake news detection, *complexity* 2021 (2021).
- [16] I. Ullah, S. Khan, M. Imran, Y.-K. Lee, Rweetminer: Automatic identification and categorization of help requests on twitter during disasters, *Expert Systems with Applications* 176 (2021) 114787.
- [17] A. M. Roy, A cnn model with feature integration for mi eeg subject classification in bmi.
- [18] A. M. Roy, An efficient multi-scale CNN model with intrinsic feature integration for motor imagery EEG subject classification in brain-machine interfaces, *Biomedical Signal Processing and Control* 74 (2022) 103496.
- [19] A. M. Roy, A multi-scale fusion cnn model based on adaptive transfer learning for multi-class mi-classification in bci system, *BioRxiv* (2022).
- [20] A. M. Roy, Adaptive transfer learning-based multiscale feature fused deep convolutional neural network for eeg mi multiclassification in brain--computer interface, *Engineering Applications of Artificial Intelligence* 116 (2022) 105347.
- [21] G. Han, J. Ma, S. Huang, L. Chen, R. Chellappa, S.-F. Chang, Multimodal few-shot object detection with meta-learning based cross-modal prompting, *arXiv preprint arXiv:2204.07841* (2022).
- [22] A. Asvadi, L. Garrote, C. Premebida, P. Peixoto, U. J. Nunes, Multimodal vehicle detection: fusing 3d-lidar and color camera data, *Pattern Recognition Letters* 115 (2018) 20--29.

- [23] S. Aleem, T. Kumar, S. Little, M. Bendechache, R. Brennan, K. McGuinness, Random data augmentation based enhancement: a generalized enhancement approach for medical datasets, arXiv preprint arXiv:2210.00824 (2022).
- [24] M. Crocco, M. Cristani, A. Trucco, V. Murino, Audio surveillance: A systematic review, ACM Computing Surveys (CSUR) 48 (4) (2016) 1--46.
- [25] Z. Fu, G. Lu, K. M. Ting, D. Zhang, A survey of audio-based music classification and annotation, IEEE transactions on multimedia 13 (2) (2010) 303--319.
- [26] W. Khan, K. Raj, T. Kumar, A. M. Roy, B. Luo, Introducing urdu digits dataset with demonstration of an efficient and robust noisy decoder-based pseudo example generator, Symmetry 14 (10) (2022) 1976.
- [27] A. M. Roy, Finite element framework for efficient design of three dimensional multicomponent composite helicopter rotor blade system, Eng 2 (1) (2021) 69--79.
- [28] W. Khan, T. Kumar, Z. Cheng, K. Raj, A. M. Roy, B. Luo, Sql and nosql databases software architectures performance analysis and assessments--a systematic literature review, arXiv preprint arXiv:2209.06977 (2022).
- [29] Y. Jiang, W. Zhang, K. Fu, Q. Zhao, Meanet: Multi-modal edge-aware network for light field salient object detection, Neurocomputing 491 (2022) 78--90.
- [30] V. A. Sindagi, Y. Zhou, O. Tuzel, Mvx-net: Multimodal voxelnet for 3d object detection, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 7276--7282.
- [31] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, K. Dietmayer, Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges, IEEE Transactions on Intelligent Transportation Systems 22 (3) (2020) 1341--1360.
- [32] A. M. Roy, R. Bose, J. Bhaduri, A fast accurate fine-grain object detection model based on YOLOv4 deep neural network, Neural Computing and Applications (2022) 1--27.

- [33] A. M. Roy, J. Bhaduri, Real-time growth stage detection model for high degree of occultation using densenet-fused YOLOv4, *Computers and Electronics in Agriculture* 193 (2022) 106694.
- [34] A. M. Roy, J. Bhaduri, A deep learning enabled multi-class plant disease detection model based on computer vision, *AI* 2 (3) (2021) 413--428.
- [35] Y. Liu, P. Sun, N. Wergeles, Y. Shang, A survey and performance evaluation of deep learning methods for small object detection, *Expert Systems with Applications* 172 (2021) 114602.
- [36] Y. Ji, H. Zhang, Z. Zhang, M. Liu, Cnn-based encoder-decoder networks for salient object detection: A comprehensive review and recent advances, *Information Sciences* 546 (2021) 835--857.
- [37] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: *Advances in neural information processing systems*, 2015, pp. 91--99.
- [38] A. Hiemann, T. Kautz, T. Zottmann, M. Hlawitschka, Enhancement of speed and accuracy trade-off for sports ball detection in videos—finding fast moving, small objects in real time, *Sensors* 21 (9) (2021) 3214.
- [39] S. K. Pal, A. Pramanik, J. Maiti, P. Mitra, Deep learning in multi-object detection and tracking: state of the art, *Applied Intelligence* 51 (9) (2021) 6400--6429.
- [40] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, B. Lee, A survey of modern deep learning based object detection models, *Digital Signal Processing* (2022) 103514.
- [41] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779--788.
- [42] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: *European conference on computer vision*, Springer, 2016, pp. 21--37.

- [43] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770--778.
- [44] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117--2125.
- [45] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, J. Sun, Detnet: A backbone network for object detection, arXiv preprint arXiv:1804.06215 (2018).
- [46] B. Jiang, R. Luo, J. Mao, T. Xiao, Y. Jiang, Acquisition of localization confidence for accurate object detection, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 784--799.
- [47] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740--755.
- [48] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International journal of computer vision 88 (2) (2010) 303--338.
- [49] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440--1448.
- [50] Z. Cai, N. Vasconcelos, Cascade r-cnn: Delving into high quality object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6154--6162.
- [51] J. R. Uijlings, K. E. Van De Sande, T. Gevers, A. W. Smeulders, Selective search for object recognition, International journal of computer vision 104 (2) (2013) 154--171.
- [52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248--255.

- [53] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, *IEEE transactions on pattern analysis and machine intelligence* 32 (9) (2010) 1627--1645.
- [54] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems* 25 (2012).
- [55] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, *IEEE transactions on pattern analysis and machine intelligence* 37 (9) (2015) 1904--1916.
- [56] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, Vol. 2, IEEE, 2006, pp. 2169--2178.
- [57] K. Lenc, A. Vedaldi, R-cnn minus r, *arXiv preprint arXiv:1506.06981* (2015).
- [58] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980--2988.
- [59] S. Liu, D. Huang, et al., Receptive field block net for accurate and fast object detection, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 385--400.
- [60] S. Zhang, L. Wen, X. Bian, Z. Lei, S. Z. Li, Single-shot refinement neural network for object detection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4203--4212.
- [61] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8759--8768.
- [62] T. Kong, F. Sun, C. Tan, H. Liu, W. Huang, Deep feature pyramid reconfiguration for object detection, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 169--185.
- [63] J. Jeong, H. Park, N. Kwak, Enhancement of ssd by concatenating feature maps for object detection, *arXiv preprint arXiv:1705.09587* (2017).

- [64] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, A. C. Berg, Dssd: Deconvolutional single shot detector, arXiv preprint arXiv:1701.06659 (2017).
- [65] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
- [66] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167 (2015).
- [67] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, D. Lin, Libra r-cnn: Towards balanced learning for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 821--830.
- [68] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, Shufflenet v2: Practical guidelines for efficient cnn architecture design, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 116--131.
- [69] J. Dai, Y. Li, K. He, J. Sun, R-fcn: Object detection via region-based fully convolutional networks, in: Advances in neural information processing systems, 2016, pp. 379--387.
- [70] Y. Zhu, C. Zhao, J. Wang, X. Zhao, Y. Wu, H. Lu, Couplenet: Coupling global structure with local parts for object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4126--4134.
- [71] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 764--773.
- [72] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, arXiv preprint arXiv:1804.02767 (2018).