

Receptive Field Block Net for Accurate and Fast Object Detection

Songtao Liu, Di Huang, Yunhong Wang

The State Key Laboratory of Virtual Reality Technology and Systems.
Beihang University, Beijing 100191, China.

{liusongtao, dhuang, yhwang}@buaa.edu.cn

Abstract

Current top-performing object detectors depend on deep CNN backbones, such as ResNet-101 and Inception, benefiting from their powerful feature representation but suffering from high computational cost. Conversely, some lightweight model based detectors fulfil real time processing, while their accuracies are often criticized. In this paper, we explore an alternative to build a fast and accurate detector by strengthening lightweight features using a crafting mechanism. Inspired by the structure of Receptive Fields (RFs) in human visual systems, we propose a novel RF Block (RFB) module, which takes the relationship between the size and eccentricity of RFs into account, to enhance the discriminability and robustness of features. We further assemble the RFB module to the top of SSD with a lightweight CNN model, constructing the RFB Net detector. To evaluate its effectiveness, experiments are conducted on two major benchmarks and the results show that RFB Net is able to reach the accuracy of advanced very deep backbone network based detectors while keeping the real-time speed. Code is available at <https://github.com/ruinmessi/RFBNet>.

1. Introduction

In recent years, Convolutional Neural Networks (CNN) have made prominent progress in object detection. Region-based CNN (R-CNN) [10], along with its representative updated descendants, *e.g.* Fast R-CNN [9] and Faster R-CNN [28], have persistently promoted detection performance on major challenges and benchmarks, such as Pascal VOC [6], MS COCO [23], and ILSVRC [29]. They formulate such an issue as a two-stage problem and build a typical pipeline, where the first phase hypothesizes category-agnostic object proposals within the given image and the second phase classifies each proposal according to CNN based deep features. It is generally accepted that CNN representation plays a crucial role, and the learned feature is expected to deliver high discriminative power encoding object related cues and good robustness especially to moderate positional shifts (usually

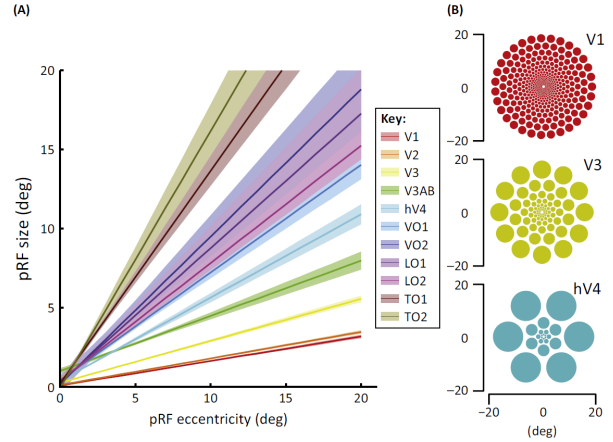


Figure 1. Regularities in human population Receptive Field (pRF) properties. (A) pRF size as a function of eccentricity in some human retinotopic maps, where two trends are evident: (1) the pRF size increases with eccentricity within each map and (2) the pRF size differs between maps. (B) The spatial array of the pRFs based on the parameters in (A). The radius of each circle is the apparent receptive field size at the appropriate eccentricity. Reproduced from [38] with the permission from J. Winawer and H. Horiguchi (<https://archive.nyu.edu/handle/2451/33887>)

incurred by inaccurate boxes). A number of very recent efforts have evidenced such a fact. For instance, [13] and [16] extract features from deeper CNN backbones, like ResNet [13] and Inception [33]; [21] introduces a top-down architecture to construct feature pyramids, integrating low-level and high-level information; and the latest top-performing Mask R-CNN [11] products an RoIAlign layer to generate more precise regional features. All these methods adopt improved feature representation to reach better results; however, advanced features basically come from deeper neural networks with heavy computational cost, making them suffer from a low inference speed.

To accelerate detection, a single-stage framework is also investigated, where the phase of object proposal generation is discarded. It examines the entire input image and applies a regular dense sampling of object locations, scales and aspect ratios, using a lightweight and faster backbone network

to extract features. Although the pioneering work, namely You Look Only Once (YOLO) [26] and Single Shot Detector (SSD) [24] illustrate the ability of real time processing, they tend to sacrifice accuracies, with a clear drop ranging from 10% to 40% relative to state-of-the-art two-stage solutions [22]. More recently, Deconvolutional SSD (DSSD) [8] and RetinaNet [22] substantially ameliorate the precision scores, which are comparable to the top ones reported by the two-stage detectors. Unfortunately such performance gains are credited to the very deep ResNet-101 [13] model as well, which limits the efficiency.

According to above discussion, to build a fast yet powerful detector, a reasonable alternative is to **enhance the feature representation of the lightweight network** by bringing in certain crafting mechanisms rather than stubbornly deepening the model. On the other side, several discoveries in neuroscience reveal that in human visual cortex, the size of population Receptive Field (pRF) is a function of eccentricity in their retinotopic maps, and although varying between maps, it increases with eccentricity within each map [38], as illustrated in Fig. 1. It helps to highlight the importance of the region nearer to the center and elevate the insensitivity to small spatial shifts. A few shallow descriptors coincidentally make use of such a mechanism to design [36] or learn [1, 39, 31] their pooling schemes, and show good performance in matching image patches.

Regarding current deep learning models, they commonly set **RFs at the same size with a regular sampling grid on a feature map, which probably induces some loss in the feature discriminability as well as robustness**. GoogleNet [34] considers RFs of multiple sizes, and it implements this concept by launching multi branch CNNs with different convolution kernels. Its variants [35, 33, 17] achieve competitive results in object detection and classification tasks. A similar idea appears in [3], where an **Atrous Spatial Pyramid Pooling (ASPP)** is exploited to capture multi-scale information and four parallel atrous convolutions with different atrous rates are applied on the top feature map. It proves effective in semantic segmentation. But the RFs of these models are several sets of concentric circles, and compared to the daisy shaped ones, the resulting feature tends to be less distinctive. Deformable CNN [4] attempts to adaptively adjust the spatial distribution of RFs according to the scale and shape of the object. Although its sampling grid is flexible, the impact of eccentricity of RFs is not taken into account, where all pixels in an RF contribute equally to the output response, making the most important information not emphasized.

Inspired by the structure of RFs in the human visual system, this paper proposes a novel module, namely Receptive Field Block (RFB), to strengthen the deep features learned from lightweight CNN models so that they can contribute to fast and accurate detectors. Specifically, RFB makes use of **multi branch pooling with varying kernels** corresponding

to RFs of **different sizes, apply dilated convolution layers to control their eccentricities**, and reshape them to generate final representation, as in Fig. 2. We then assemble the RFB module to the top of SSD [24], a real time approach with a lightweight backbone, and construct the advanced one-stage detector (RFB Net). Thanks to such a simple module, RFB Net delivers relatively decent scores that are comparable to the ones of up-to-date deeper backbone network based detectors [21, 20, 22] and retains the real-time speed of the origin lightweight detector. Additionally, the RFB module is more generic and imposes less constraints on the network architecture.

Our main contributions can be summarized as follows:

1. We propose the RFB module to simulate the configuration in terms of the size and eccentricity of RFs in human visual systems, aiming to enhance deep features of lightweight CNN networks.
2. We present the RFB Net based detector, and by simply replacing the top convolution layers of SSD [24] with RFB, it shows significant performance gain while still keeping the computational cost under control.
3. We show that RFB Net achieves state-of-the-art results on Pascal VOC and MS COCO at a real time processing speed, and demonstrate the generalization ability of RFB by linking it to MobileNet [14].

2. Related Work

Classic detector: The traditional approaches to object detection are either sliding window based or region proposal based. Deformable Part Model (DPM) [7] as well as Selective Search (SS) [37] are two milestones, and they both show remarkable performance at one time. Since the basic deep learning model, *i.e.* AlexNet [18], achieved the breathtaking improvement in image classification, two-stage detectors, described next, have quickly dominated this field.

Two-stage detector: R-CNN [10] straightforwardly combines the steps of cropping box proposals like SS and classifying them through a CNN model, yielding a significant accuracy gain, which opens the deep learning era. R-CNN is computationally expensive as it has to judge thousands of image patches. For speeding up, Fast R-CNN [9] computes the entire image only once in a feature extractor and then puts it into a spatial pooling layer, called ROI pooling, thus allowing to reuse the features in classification.

[28, 41] show that the quality of object proposals can be optimized by deep neural networks, and Faster R-CNN [28] replaces the independent proposal generators in its predecessors by Region Proposal Network (RPN). RPN has a set of boxes, named anchors, paved on the image at different locations, scales and aspect ratios, and it is trained to make a class-agnostic prediction and a regression prediction of an

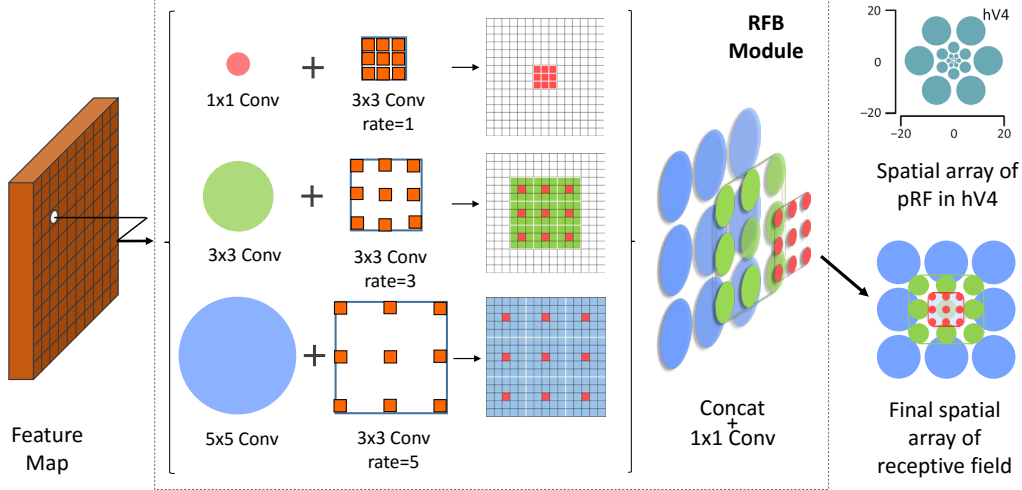


Figure 2. Construction of the RFB module by combining multiple branches with different kernels and dilated convolution layers. Multiple kernels are analogous to the pRFs of varying sizes, while dilated convolution layers assign each branch with an individual eccentricity to simulate the ratio between the size and eccentricity of the pRF. With a concatenation and 1×1 conv in all the branches, the final spatial array of RF is produced, which is similar to that in human visual systems, as depicted in Fig. 1.

offset fit the object location for each anchor. Such a framework is later extended to many more advanced versions. Although Faster R-CNN runs much faster than Fast R-CNN does, it still needs to apply region-specific computation for hundreds of times.

R-FCN [19] is a fully-convolutional variant that greatly reduces the computational cost on each proposal by removing fully-connected layers and adopting position-sensitive score maps for the final prediction. Another recent extension of Faster R-CNN is Mask R-CNN [11], which adds a parallel branch to segment the object mask and employs the RoIAlign layer to fix misalignment, further improving the detection accuracy.

One-stage detector: The most representative one-stage detectors are YOLO [26, 27] and SSD [24]. YOLO predicts confidences and locations for multiple objects based on the whole feature map. It runs at a high speed by eliminating the stage of proposal generation, but it struggles in precisely localizing some objects, especially the small ones. SSD can be deemed as a multi-scale version of YOLO in some sense. It makes use of multi-scale feature maps and RPN-like default anchor boxes to construct a more accurate but still fast detector. These detectors both adopt lightweight backbones to speed up detection, while their accuracies apparently trail those of top two-stage methods.

An advanced version of SSD, called DSSD [8], replaces the original lightweight backbone to the deeper ResNet-101 [13] and adds a deconvolution module to improve the quality of deep features, which reports a better result. Another recent single-stage detector, named RetinaNet [22], is also built based on ResNet-101. It applies a novel Focal Loss to handle class imbalance and easy negatives dominating, and outperforms all existing state-of-the-art two-stage methods.

However, such performance gain largely consumes their advantage in speed.

Receptive field: Recall that in this study, we aim to improve the detection performance of high-speed single-stage detectors without incurring too much computational burden. Therefore, instead of applying deeper backbones, RFB, imitating the mechanism of RFs in the human visual system, is used to enhance lightweight model based feature representation. Actually, there exist several studies that discuss RFs in CNN, and the most related ones are GoogleNet and its variants [34, 35, 33], Dilated Convolution [3], and Deformable CNN [4]. Together with RFB, all these methods capture information through multi-scale RFs implemented in different manners, and share the properties of end-to-end learning, convenient training, and easy integrating into any CNN architectures. Nevertheless, in contrast to the three counterparts, RFB is biological vision inspired, which emphasizes the relationship between RF size and eccentricity in a daisy-shape configuration. See Fig. 3 for differences of the three typical spatial RF structures.

3. Method

In this section, we revisit the human visual cortex, introduce our RFB components and the way to simulate such a mechanism, and describe the RFB Net detector architecture and the training/testing schedule.

3.1. Visual Cortex Revisit

During the past few decades, it has come true that functional Magnetic Resonance Imaging (fMRI) non-invasively measures human brain activities at a resolution in millimeter, and RF modeling has become an important sensory sci-

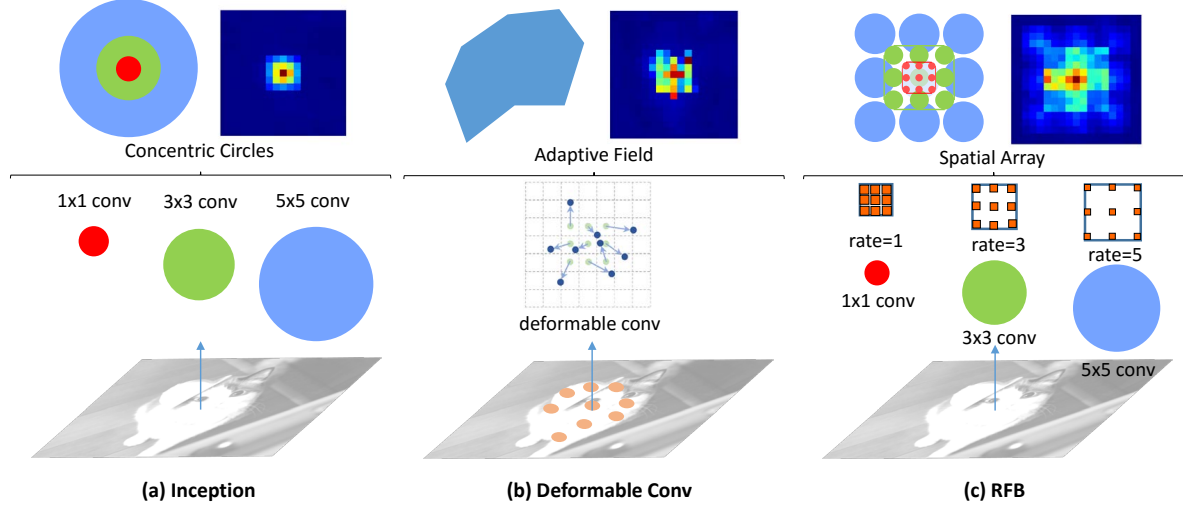


Figure 3. Three typical structures of Spatial RFs. (a) shows the kernels of multiple sizes in the Inception module, and the final RF concat forms a set of concentric circles. (b) adopts deformable conv to produce an adaptive RF according to the object characteristics. (c) illustrates the mechanism of our RFB module. The color map of each structure is the effective RF derived from one correspondent layer in the trained model, depicted by using the same gradient back-propagation method in [25].

ence tool used to predict responses and clarify brain computations. Since human neuroscience instruments often observe the pooled responses of many neurons, these models are thus commonly called pRF models [38]. Based on fMRI and pRF modeling, it is possible to investigate the relation across many visual field maps in the cortex. At each cortical map, researchers find a positive correlation between pRF size and eccentricity [38], while the coefficient of correlation varies between visual field maps, as shown in Fig. 1.

3.2. Receptive Field Block

The proposed RFB is a multi-branch convolutional block similar to the Inception [34] block. Its inner structure can be divided into two components: the multi-branch convolution layer with different kernels and the trailing dilated pooling or convolution layers. The former part is identical to that of Inception, responsible for simulating the pRFs of multiple sizes, and the latter part reproduces the relation between the pRF size and eccentricity in the human visual system. Fig. 2 illustrates this proposed RFB along with its corresponding spatial pooling region maps. We elaborate the two parts and their functions in detail in the following.

Multi-branch convolution layer: According to the definition of RF in CNNs, it is a simple and natural way to apply different kernels to realize RFs of multiple sizes, which is supposed to be superior to that where RFs share a fixed size. The Inception series [34, 35, 33] clearly demonstrate the effectiveness of this construction in object detection and image classification [16]. In addition, it is also evidenced in our case, the accuracy of SSD is improved by replacing each top convolution layer with an Inception block (see Table 3).

We adopt the latest changes in the updated versions, *i.e.*, Inception V4 and Inception-ResNet V2 [33] in the Inception

family. To be specific, first, we employ the bottleneck structure in each branch, consisting of a 1×1 conv-layer to decrease the number of channels in feature map plus an $n \times n$ conv-layer. Second, we replace the 5×5 conv-layer by two stacked 3×3 conv-layers to reduce parameters and deeper non-linear layers. For the same reason, we use a $1 \times n$ plus an $n \times 1$ conv-layer to take place of the original $n \times n$ conv-layer. Ultimately, we apply the shortcut design from ResNet [13] and Inception-ResNet V2 [33]. Since the top convolution feature layers often have 2 stride or decreased feature maps in output, we change the shortcut from identity mapping to a 1×1 conv-layer without non-linear activation.

Dilated pooling or convolution layer: This concept is first introduced in Deeplab [2], which is also named the astrous convolution layer. The original intention of this structure is to replace the 2-stride convolution layer to generate feature maps of a higher resolution, capturing information at a larger context while keeping the same number of parameters and the same RF. This design has rapidly proved competent at semantic segmentation [3], and has also been adopted in some reputable object detectors, such as SSD [24] and R-FCN [19], to elevate speed or/and accuracy.

In this paper, we introduce dilation convolution to simulate the impact of the eccentricities of pRFs in human visual cortex. Fig. 4 and Fig. 5 illustrate two combinations of multi-branch convolution layer and dilated pooling or convolution layer. At each branch, the convolution layer of a particular kernel size is followed by a pooling or convolution layer with corresponding dilation. The kernel size and dilation have a similar positive functional relation as that of the size and eccentricity of pRFs in the visual cortex. Eventually, the feature maps of all the branches are concatenated together, merging into a spatial pooling or convolution ar-

ray as in Fig. 1. When using a max or an average pooling layer with dilation, the RFB contains less parameters, but it loses the flexibility in reorganizing the features from RFs of different sizes. While choosing a dilated convolution layer, RFB enhances the feature representation power since the final feature map can be seen as a learnable linear combination of the spatial array, although its number of parameters moderately increases. We show the trade-off in the selection step in Section 4.2

The specific parameters of RFB, like kernel size or dilation of each branch and the number of branches are slightly different at each position within the detector, which will be clarified in the next section.

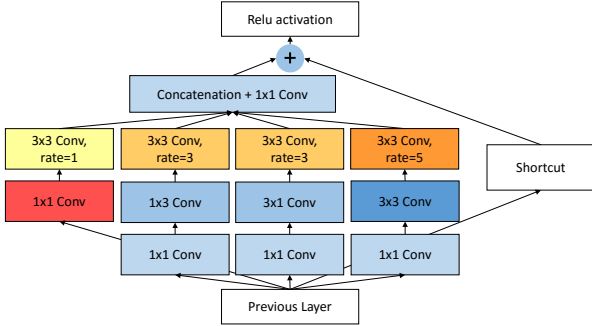


Figure 4. The architecture of RFB-a. This module is employed to mimic smaller pRFs in shallow human retinotopic maps, using more branches with smaller kernels.

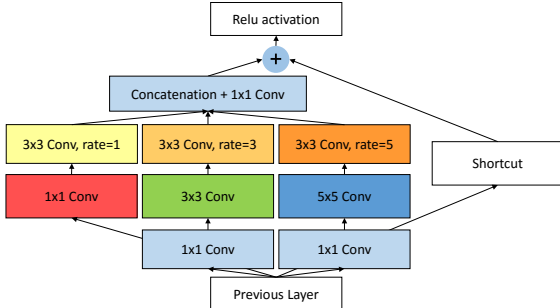


Figure 5. The architecture of RFB-b. Following [35], we use two layers of 3×3 conv replacing 5×5 to reduce parameters, which is not shown for better visualization.

3.3. RFB Net Detection Architecture

The proposed RFB Net detector reuses the multi-scale, one-stage detection framework of SSD [24], where the proposed RFB module is embedded to ameliorate the feature extracted from the lightweight backbone so that the detector is more accurate and fast enough. Thanks to the property of RFB for easy integrating into CNNs, we can preserve the SSD architecture as much as possible. The main modification lies in replacing top convolution layers with RFB, and some minor but active ones are given in Fig. 6. We describe each component in the RFB Net detector next.

Lightweight backbone: We use exactly the same backbone network as in SSD [24]. In brief, it is a VGG16 [32] architecture pre-trained on the ILSVRC CLS-LOC dataset [29], where its fc6 and fc7 layers are converted to convolutional layers with sub-sampling parameters, and its pool5 layer is changed from 2×2 -s2 to 3×3 -s1. The dilated convolution layer is used to fill holes and all the dropout layers and the fc8 layer are removed. Even though many accomplished lightweight networks have recently been proposed (e.g. DarkNet [27], MobileNet [14], and ShuffleNet [42]), we focus on this backbone to achieve direct comparison to the original SSD [24].

RFB on multi-scale feature maps: In the original SSD [24], the base network is followed by a cascade of convolutional layers to form a series of feature maps with consecutively decreasing spatial resolutions and increasing fields of view. We keep the same cascade structure, but the front convolutional layers with feature maps of relatively large resolutions are replaced by the RFB-b module and the conv4_3 features are also tailed an RFB-b layer instead of L2 normalization. In the primary version of RFB, we use a dilated max pooling to imitate the impact of eccentricity. The last few convolutional layers are preserved since the resolutions of their feature maps are too small to apply filters with large kernels like 5×5 . This simple substitution upgrades the original SSD (see Table 2).

Tuning RFB parameters: As exhibited in Fig. 1, the rate of the size and eccentricity of pRF differs between visual maps. This variation is naturally adapted in our RFB by correspondingly tuning the kernel size and dilation as illustrated in Fig. 4. We thus adjust the parameters of the RFB to form an RFB-a module behind the conv4_3, to add more small filters for possible improvements (see Table 2).

3.4. Training and Inference Settings

We implement our RFB Net detector based on the framework of Pytorch¹, utilizing several parts of open source infrastructures provided by the `ssd.pytorch` repository². Our training strategies mostly follow SSD, including data augmentation, hard negative mining, scale and aspect ratios for default boxes and loss function (e.g., smooth L1 loss for localization and softmax loss for classification), while we slightly change our learning rate scheduling for better accommodation of RFB. More details are given in the following section of experiments. All new conv-layers are initialized with the MSRA method [12].

With a potentially large number of boxes generated from our detector, performing non-maximum suppression (NMS) might significantly slow down the inference. We thus suggest a more efficient post-processing strategy when there are too many proposals. We first use a confidence threshold of

¹<https://pytorch.org/>

²<https://github.com/amdegroot/ssd.pytorch>

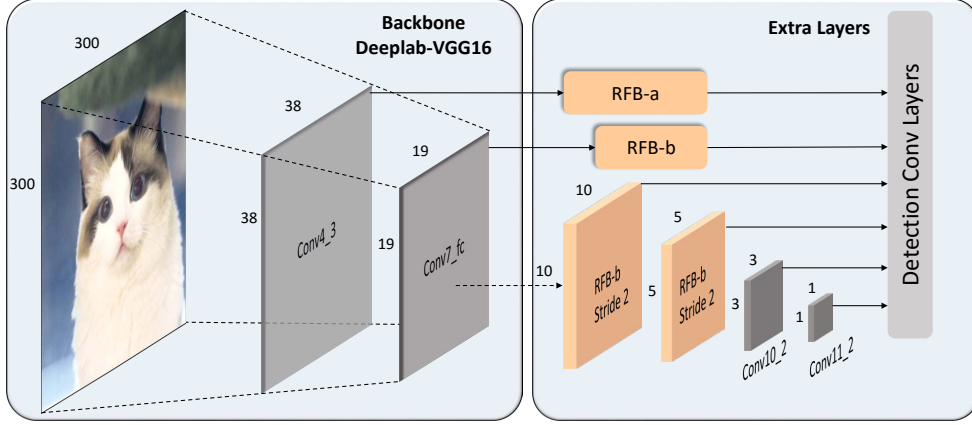


Figure 6. The pipeline of RFB-Net300. The conv4_3 feature map tails the RFB-a module which has smaller RFs and an RFB-b module with stride 2 is produced by operating 2 stride of multi-kernel conv-layers in the original RFB-b.

0.015 to filter out most boxes, which is a little higher than 0.01 in SSD, and we then pre-select the top 200 boxes with the largest scores and apply NMS with jaccard overlap of 0.45 for each class, leaving only 50 per class. Overall, final detection is conducted on the top 200 boxes per image. This step notably shrinks the computational time for NMS, especially in the COCO dataset with 80 classes, and has a marginal impact on accuracy.

4. Experiments

We conduct experiments on the MS COCO [23] and Pascal VOC 2007 datasets [6], which have 80 and 20 object categories respectively. In the VOC 2007 database, a predicted bounding box is positive if its Intersection over Union (IoU) with the ground truth is higher than 0.5, while in COCO, it uses various thresholds for a more comprehensive calculation. The metric for evaluating detection performance is the mean Average Precision (mAP). Code is available at <https://github.com/ruinmessi/RFBNet>.

4.1. Pascal VOC 2007

In this experiment, we train our RFB Net on the union of 2007 *trainval* set and 2012 *trainval* set. We set the batch size at 32 and the initial learning rate of 10^{-3} as in the original SSD model, but it makes the training process not so stable as the loss drastically fluctuates. Instead, we use a “warmup” strategy that gradually ramps up the learning rate from 10^{-6} to 4×10^{-3} at the first 5 epochs. After the “warmup” phase, it goes back to the original learning rate schedule divided by 10 at 150 and 200 epochs. The total number of training epochs is 250. Following [24], we utilize a weight decay of 0.0005 and a momentum of 0.9.

Table 1 shows the comparison between our results and the state of the art ones on the VOC2007 *test* set. SSD300* and SSD512* are the updated SSD results with the new expansion of the data augmentation technique. By integrating

the RFB layers, our basic model, *i.e.* RFB Net300, outperforms SSD and YOLO with an mAP of 80.5%, while almost keeping the same speed as SSD300. It even reaches the same accuracy with R-FCN [19], the advanced model under the two stage framework. RFB Net512 performs 82.2% mAP with a larger input size, better than most object detection systems for both one stage and two stage equipped with very deep base backbone networks, while still running at a high speed.

Method	Backbone	Data	mAP(%)	FPS
Faster [28]	VGG	07+12	73.2	7
Faster [13]	ResNet-101	07+12	76.4	5
R-FCN [19]	ResNet-101	07+12	80.5	9
YOLOv2 544 [27]	Darknet	07+12	78.6	40
R-FCN w Deformable CNN [4]	ResNet-101	07+12	82.6	8*
SSD300* [24]	VGG	07+12	77.2	46
DSSD321 [8]	ResNet-101	07+12	78.6	9.5
RFB Net300	VGG	07+12	80.5	43
SSD512* [24]	VGG	07+12	79.8	19
DSSD513 [8]	ResNet-101	07+12	81.5	5.5
RFB Net512	VGG	07+12	82.2	18

* Extrapolated time

Table 1. Comparison of detection methods on the PASCAL VOC 2007 *test* set. All runtime information is computed on a Graphics card of Geforce GTX Titan X (Maxwell architecture).

4.2. Ablation on VOC 2007

RFB module: For better understanding RFB, we investigate the impact of each component in its design, and also compare RFB with some similar structures. The results are summarized in Table 2 and Table 3. As displayed in Table 2, the original SSD300 with new data augmentation achieves 77.2% mAP. By simply replacing the last convolution layer with the RFB-max pooling, we can see that the result is improved to 79.1%, delivering a gain of 1.9% in mAP, which indicates that the RFB module is effective in detection.

Cortex map simulation: As described in Sec.3.3, we tune our RFB parameters to simulate the ratio between the size and eccentricity of pRFs in cortex maps. This adjust-

ment boosts the performance by 0.5%, which validates the mechanism in human visual systems (Table 2).

More prior anchors: The original SSD associates only 4 default boxes at conv4_3, conv10_2, and conv11_2 feature map locations and 6 default anchors for all the other layers. Recent research [15] claims that low level features are critical to detecting small objects. We thus suppose that performance, especially that of small instances, tends to increase if more anchors are added in low level feature maps like conv4_3. In the experiment, we put 6 default priors at conv4_3, and it has no influence on the original SSD, while it further improves 0.2% mAP for our RFB model (Table 2).

Dilated convolutional layer: In early experiments, we choose dilated pooling layers for our RFB module to avoid incurring additional parameters, but these stationary pooling strategies limit feature fusion of RFs of multiple sizes. When picking the dilated convolutional layer, we find that it raises the accuracy by 0.7% mAP without slowing down the inference speed (Table 2).

Block architecture: We also compare our RFB to Deformable CNN [4], Inception [34], Dilated Convolution [3], ResNet [13], ResNext [40], and several RFB-like modules with special settings. For the RFB-like modules, “fixed eccentricity” means all the dilated convolution layers at multi branches have the same dilation at 1; “fixed RF” sets the kernel size of all the convolution layers in the module to 3×3 ; and “negative ratio” inverses the ratio of kernel size and dilation. We keep all the different structures have the same training schedule and almost the same number of parameters. Their evaluations are recorded in Table 3, and we can see that our RFB performs best. Specially, RFB surpasses Dilated Conv with rate at 8, which has the same size of entire RF. It points out that the dedicated RFB structure indeed contributes to the detection precision.

	SSD*					RFB
RFB-max pooling?	✓	✓	✓			
Tuning RFB?		✓	✓	✓	✓	✓
More Prior?			✓	✓	✓	✓
RFB-avg pooling?				✓		
RFB-dilated conv?						✓
	77.2	79.1	79.6	79.8	79.8	80.5

Table 2. Effectiveness of various designs on the VOC 2007 *test* set. Refer to Section 3.3 and Section 4.2 for more details

4.3. Microsoft COCO

To further validate the proposed RFB module, we carry out experiments on the MS COCO dataset. Following [24, 22], we use the *trainval35k* set (*train* set + *val* 35k set) for training and set the batch size at 32. We keep the original SSD strategy that decreases the size of default boxes, since objects in COCO tend to be smaller than those in PASCAL VOC. At the begin of training, we still apply the “warmup” technique that progressively increases the learning rate from

Architecture	#parameters	mAP (%)
RFB	34.3M	80.5
Fixed Eccentricity	34.3M	79.2
Fixed RF	33.7M	79.7
Negative Ratio	34.3M	79.2
Deformable CNN [4]	35.2M	79.5
Dilated Conv (Rate=8) [3]	38.3M	79.1
Inception [34]	33.9M	78.9
ResNet [13]	34.5M	78.4
ResNext [40]	33.9M	78.6

Table 3. Performance comparison of different block architectures on the VOC 2007 *test* set.

10^{-6} to 2×10^{-3} at the first 5 epochs, then decrease it after 80 and 100 epochs by the factor of 10, and end up at 120.

From Table 4, it can be seen that RFB Net300 achieves 29.9%/49.9% on the *test-dev* set, which surpasses the baseline score of SSD300* with a large margin, and even equals to the R-FCN [19] which employs ResNet-101 as the base net with a larger input size (600×1000) under the two stage framework.

Regarding the bigger model, the result of RFB Net512 is slightly inferior to but still comparable to the one of the recent advanced one-stage model RetinaNet500 (33.8% vs. 34.4%). However, it should be noted that RetinaNet makes use of the deep ResNet-101-FPN backbone and a new loss to make learning focus on hard examples, while our RFB Net is only built on a lightweight VGG model. On the other hand, we can see that RFB Net500 averagely consumes 55 ms per image, while RetinaNet needs 90 ms.

One may notice that RetinaNet800 [22] reports the top accuracy (39.1%) based on a very high image resolution up to 800 pixels. Although it is well known that a larger input image size commonly yields higher performance, it is out of the scope of this study, where accurate and fast detector is pursued. Instead, we consider two efficient updates: (1) to up-sample the conv7_fc feature maps and concat it with the conv4_3 before applying the RFB-a module, sharing a similar strategy as in FPN [21]; and (2) to add a branch with a 7×7 kernel in all RFB layers. As we can see in Table 4, they further increase the performance, making the best score in this study at 34.4% (denoted as RFB Net512-E), while the computational cost marginally ascends.

5. Discussion

Inference speed comparison: In Table 1 and Fig. 7, we show speed comparison to other recent top-performing detectors. In our experiments, the inference speeds in different datasets have slight variations, since MS COCO has 80 categories and average dense instances consume more time on the NMS process. Table 1 shows that our RFB Net300 is the most accurate one (80.5% mAP) among the real time detectors and runs at 43 fps in the Pascal VOC dataset, and RFB Net512 provides more accurate detections even with a speed close to real time (18 fps). In Fig. 7, we follow [22]

Method	Backbone	Data	Time	Avg. Precision, IoU:			Avg. Precision, Area:		
				0.5:0.95	0.5	0.75	S	M	L
Faster [28]	VGG	trainval	147 ms	24.2	45.3	23.5	7.7	26.4	37.1
Faster+++ [13]	ResNet-101	trainval	3.36 s	34.9	55.7	37.4	15.6	38.7	50.9
Faster w FPN [21]	ResNet-101-FPN	trainval35k	240 ms	36.2	59.1	39.0	18.2	39.0	48.2
Faster by G-RMI [16]	Inception-Resnet-v2 [33]	trainval	–	34.7	55.5	36.7	13.5	38.1	52.0
R-FCN [19]	ResNet-101	trainval	110 ms	29.9	51.9	–	10.8	32.8	45.0
R-FCN w Deformable CNN [4]	ResNet-101	trainval	125 ms*	34.5	55.0	–	14.0	37.7	50.3
Mask R-CNN [11]	ResNext-101-FPN	trainval35k	210 ms	37.1	60.0	39.4	16.9	39.9	53.5
YOLOv2 [27]	darknet	trainval35k	25 ms	21.6	44.0	19.2	5.0	22.4	35.5
SSD300* [24]	VGG	trainval35k	22 ms	25.1	43.1	25.8	–	–	–
SSD512* [24]	VGG	trainval35k	53 ms	28.8	48.5	30.3	–	–	–
DSSD513 [8]	ResNet-101	trainval35k	182 ms	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet500 [22]	ResNet-101-FPN	trainval35k	90 ms	34.4	53.1	36.8	14.7	38.5	49.1
RetinaNet800 [22]	ResNet-101-FPN	trainval35k	198 ms	39.1	59.1	42.3	21.8	42.7	50.2
RFB Net300	VGG	trainval35k	25 ms	29.9	49.9	31.1	11.9	31.8	44.7
RFB Net512	VGG	trainval35k	55 ms	33.8	54.2	35.9	16.2	37.1	47.4
RFB Net512-E	VGG	trainval35k	59 ms	34.4	55.7	36.4	17.6	37.0	47.6

* Extrapolated time

Table 4. Detection performance on the COCO *test-dev* 2015 dataset. Almost all the methods are measured on the Nvidia Titan X (Maxwell architecture) GPU, except RetinaNet, Mask R-CNN and FPN (Nvidia M40 GPU).

to plot the speed/accuracy trade-off curve for RFB Net, and compare it to RetinaNet [22] and other recent methods on the MS COCO *test-dev* set. This plot displays that our RFB Net forms an upper envelope among all the real time detectors. In particular, RFB Net300 keeps a high speed (40 fps) while outperforming all the high frame rate counterparts. Note that they are measured on the same Titan X (Maxwell architecture) GPU, except RetinaNet (Nvidia M40 GPU).

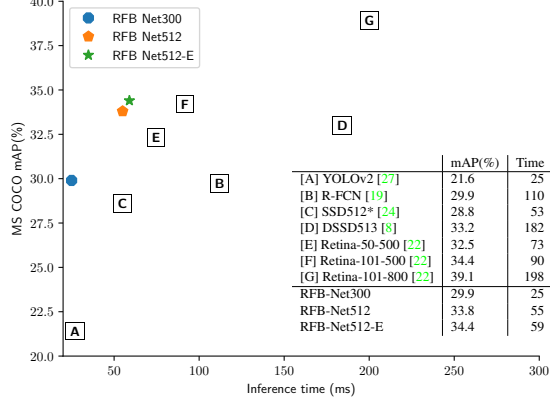


Figure 7. Speed (ms) vs. accuracy (mAP) on MS COCO *test-dev*. Enabled by the proposed RFB module, our single one-stage detector surpasses all existing high frame rate detectors, including the best reported one-stage system Retina-50-500 [22].

Other lightweight backbone: Although the base backbone we use is a reduced VGG16 version, it still has a large number of parameters compared with those recent advanced lightweight networks, *e.g.*, MobileNet [14], DarkNet [27], and ShuffleNet [42]. To further test the generalization ability of the RFB module, we link RFB to the MobileNet-SSD version [14]. Following [14], we train it on the MS COCO *train+val35k* dataset with the same schedule and make evaluation on *minival*. Table 5 shows that RFB still increases the accuracy of the MobileNet backbone with limited additional

Framework	Model	mAP(%)	#parameters
SSD 300	MobileNet [14]	19.3%	6.8M
SSD 300	MobileNet+RFB	20.5%	7.4M

Table 5. Accuracies on MS COCO *minival2014* using MobileNet as the backbone.

layers and parameters. This suggests its great potential for applications on low-end devices.

Training from scratch: We also notice another interesting property of the proposed RFB module, *i.e.* efficient in training the object detector from scratch. Recently, according to [30], training without using pre-trained backbones is discovered to be a hard task, where all the structures of base nets fail to train from scratch in the two-stage framework and the prevalent CNNs (ResNet or VGG) in the one-stage framework successfully converge with much worse results. Deeply Supervised Object Detectors (DSOD) [30] proposes a lightweight structure which achieves 77.7% mAP on the VOC 2007 *test* set without pre-training, but it fails to promote the performance when using pre-trained network. We train our RFB Net300 on the VOC 07+12 *trainval* set from scratch and reach 77.6% mAP on the same test set, which is comparable to DSOD. It is worth noting that our pre-trained version can boost the performance to 80.5% mAP.

6. Conclusion

In this paper, we propose a fast yet powerful object detector. In contrast to the widely employed way that greatly deepens the backbone, we choose to enhance feature representation of lightweight networks by bringing in a crafting mechanisms, namely Receptive Field Block (RFB), which imitates the structure of RF in human visual systems. RFB measures the relationship between the size and eccentricity of RFs, and generates more discriminative and robust features. RFB is equipped on the top of lightweight CNN

based SSD, and the resulting detector delivers a significant performance gain on the Pascal VOC 2007 and MS COCO databases, where the final accuracies are even comparable to those of existing top-performing deeper model based detectors. In addition, it retains the advantage in processing speed of lightweight models.

7. Pascal VOC 2012

We also evaluate our RFB Net on the Pascal VOC2012 dataset. In this experiment, we train our RFB Net on the union of 2007 *trainval*+ *test* set and 2012 *trainval* set, keeping the same training strategy as on VOC2007.

As some top-performing models have not reported their results on VOC2012, *e.g.*, Mask RCNN [11], RetinaNet [22], we add two advanced methods for comparison, which are the state of the arts achieved by single models on the VOC2012 *test* set. BlitzNet [5] is a recent deep system for object detection and segmentation, using both bounding boxes and segmentation labels in training. CoupleNet [43] adopts local information partly based on R-FCN [19]. Table 6 shows that our RFB Net obtains a top mAP of 81.2%, which is 0.8 points higher than CoupleNet [43]. We note that without using extra data and extra tricks in the testing phase, our RFB Net512 is the first one with an mAP higher than 81%, while still running at a high speed.

References

- [1] M. Brown, G. Hua, and S. Winder. Discriminative learning of local image descriptors. *TPAMI*, 2011. 2
- [2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016. 4
- [3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2, 3, 4, 7
- [4] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *ICCV*, 2017. 2, 3, 6, 7, 8
- [5] N. Dvornik, K. Shmelkov, J. Mairal, and C. Schmid. BlitzNet: A real-time deep network for scene understanding. In *ICCV*, 2017. 9, 10
- [6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 1, 6
- [7] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 2
- [8] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017. 2, 3, 6, 8, 10
- [9] R. Girshick. Fast r-cnn. In *ICCV*, 2015. 1, 2
- [10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1, 2
- [11] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 3, 8, 9
- [12] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 5
- [13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 2, 3, 4, 6, 7, 8, 10
- [14] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2, 5, 8
- [15] P. Hu and D. Ramanan. Finding tiny faces. In *CVPR*, 2017. 7
- [16] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *CVPR*, 2017. 1, 4, 8
- [17] K.-H. Kim, S. Hong, B. Roh, Y. Cheon, and M. Park. Pvanet: Deep but lightweight neural networks for real-time object detection. *arXiv preprint arXiv:1608.08021*, 2016. 2
- [18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2
- [19] Y. Li, K. He, J. Sun, et al. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, 2016. 3, 4, 6, 7, 8, 9, 10
- [20] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In *CVPR*, 2017. 2
- [21] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1, 2, 7, 8
- [22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2, 3, 7, 8, 9
- [23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 6
- [24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 2, 3, 4, 5, 6, 7, 8, 10
- [25] W. Luo, Y. Li, R. Urtasun, and R. Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *NIPS*, 2016. 4
- [26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 2, 3
- [27] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In *CVPR*, 2017. 3, 5, 6, 8, 10
- [28] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1, 2, 6, 8
- [29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 1, 5
- [30] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue. Dsod: Learning deeply supervised object detectors from scratch. In *ICCV*, 2017. 8

network	data	backbone	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv
Faster [13]	07++12	ResNet-101	73.8	86.5	81.6	77.2	58.0	51.0	78.6	76.6	93.2	48.6	80.4	59.0	92.1	85.3	84.8	80.7	48.1	77.3	66.5	84.7	65.6
R-FCN [19]	07++12	ResNet-101	77.6	86.9	83.4	81.5	63.8	62.4	81.6	81.1	93.1	58.0	83.8	60.8	92.7	86.0	84.6	84.4	59.0	80.8	68.6	86.1	72.9
CoupleNet [43]	07++12	ResNet-101	80.4	89.1	86.7	81.6	71.0	64.4	83.7	83.7	94.0	62.2	84.6	65.6	92.7	89.1	87.3	87.7	64.3	84.1	72.5	88.4	75.3
YOLOv2 [27]	07++12	Darknet	73.4	86.3	82.0	74.8	59.2	51.8	79.8	76.5	90.6	52.1	78.2	58.5	89.3	82.5	83.4	81.3	49.1	77.2	62.4	83.8	68.7
SSD300* [24]	07++12	VGG	75.8	88.1	82.9	74.4	61.9	47.6	82.7	78.8	91.5	58.1	80.0	64.1	89.4	85.7	85.5	82.6	50.2	79.8	73.6	86.6	72.1
DSSD321 [8]	07++12	ResNet-101	76.3	87.3	83.3	75.4	64.6	46.8	82.7	76.5	92.9	59.5	78.3	64.3	91.5	86.6	86.6	82.1	53.3	79.6	75.7	85.2	73.9
BlitzNet300 [5]	07+12+S	ResNet50	75.4	87.4	82.1	74.5	61.6	45.9	81.5	78.3	91.4	58.2	80.3	64.9	89.1	83.5	85.7	81.5	50.5	79.9	74.7	84.8	71.1
RFBNet300	07++12	VGG	78.7[†]	89.9	86.8	76.1	65.3	54.8	85.2	81.9	92.1	62.5	83.9	65.9	90.9	87.6	88.2	85.1	55.9	83.5	76.2	87.3	74.9
SSD512* [24]	07++12	VGG	78.5	90.0	85.3	77.7	64.3	58.5	85.1	84.3	92.6	61.3	83.4	65.1	89.9	88.5	88.2	85.5	54.4	82.4	70.7	87.1	75.6
DSSD513 [8]	07++12	ResNet-101	80.0	92.1	86.6	80.3	68.7	58.2	84.3	85.0	94.6	63.3	85.9	65.6	93.0	88.5	87.8	86.4	57.4	85.2	73.4	87.8	76.8
BlitzNet512 [5]	07+12+S	ResNet50	79.0	89.9	85.2	80.4	67.2	53.6	82.9	83.6	93.8	62.5	84.0	65.8	91.6	86.6	87.6	84.6	56.8	84.7	73.9	88.0	75.7
RFBNet512	07++12	VGG	81.2[‡]	90.9	89.1	80.0	70.1	64.6	87.1	86.8	93.1	64.7	85.4	68.3	91.2	89.4	88.3	87.6	60.4	87.4	74.2	88.2	77.4

Table 6. **Detection results on PASCAL VOC 2012.** For fair comparison, we only list the results of single models without bells and whistles in the testing phase. “07++12”: 07 *trainval* + 07 *test* + 12 *trainval*, “07+12+S”: 07 + 12 *trainval* plus segmentation labels. †: <http://host.robots.ox.ac.uk:8080/anonymous/CQ9HCJ.html>, ‡: <http://host.robots.ox.ac.uk:8080/anonymous/XF4XT6.html>

- [31] K. Simonyan, A. Vedaldi, and A. Zisserman. Learning local feature descriptors using convex optimisation. *TPAMI*, 2014. **2**
- [32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *NIPS*, 2014. **5**
- [33] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017. **1, 2, 3, 4, 8**
- [34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. **2, 3, 4, 7**
- [35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. **2, 3, 4, 5**
- [36] E. Tola, V. Lepetit, and P. Fua. A fast local descriptor for dense matching. In *CVPR*, 2008. **2**
- [37] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *IJCV*, 2013. **2**
- [38] B. A. Wandell and J. Winawer. Computational neuroimaging and population receptive fields. *Trends in cognitive sciences*, 2015. **1, 2, 4**
- [39] S. A. Winder and M. Brown. Learning local image descriptors. In *CVPR*, 2007. **2**
- [40] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. **7**
- [41] S. Zagoruyko, A. Lerer, T.-Y. Lin, P. O. Pinheiro, S. Gross, S. Chintala, and P. Dollár. A multipath network for object detection. *arXiv preprint arXiv:1604.02135*, 2016. **2**
- [42] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *arXiv preprint arXiv:1707.01083*, 2017. **5, 8**
- [43] Y. Zhu, C. Zhao, J. Wang, X. Zhao, Y. Wu, and H. Lu. Couplenet: Coupling global structure with local parts for object detection. In *ICCV*, 2017. **9, 10**