

# Rethinking Classification and Localization for Object Detection

Yue Wu<sup>1</sup>, Yinpeng Chen<sup>2</sup>, Lu Yuan<sup>2</sup>, Zicheng Liu<sup>2</sup>, Lijuan Wang<sup>2</sup>, Hongzhi Li<sup>2</sup> and Yun Fu<sup>1</sup>  
<sup>1</sup>Northeastern University, <sup>2</sup>Microsoft

{yuewu, yunfu}@ece.neu.edu, {yiche, luyuan, zliu, lijuanw, hongzhi.li}@microsoft.com

## Abstract

Two head structures (i.e. fully connected head and convolution head) have been widely used in R-CNN based detectors for classification and localization tasks. However, there is a lack of **understanding** of how does these two head structures work for these two tasks. To address this issue, we perform a thorough analysis and find an interesting fact that the two head structures have **opposite** preferences towards the two tasks. Specifically, the fully connected head (fc-head) is more suitable for the classification task, while the convolution head (conv-head) is more suitable for the localization task. Furthermore, we examine the output feature maps of both heads and find that fc-head has more spatial sensitivity than conv-head. Thus, fc-head has more capability to distinguish a complete object from part of an object, but is not robust to regress the whole object. Based upon these findings, we propose a **Double-Head** method, which has a fully connected head focusing on classification and a convolution head for bounding box regression. Without bells and whistles, our method gains +3.5 and +2.8 AP on MS COCO dataset from Feature Pyramid Network (FPN) baselines with ResNet-50 and ResNet-101 backbones, respectively.

## 1. Introduction

Most two-stage object detectors [10, 11, 35, 4, 26] share a head for both classification and bounding box regression. Two different head structures are widely used. Faster R-CNN [35] uses a convolution head (conv5) on a single level feature map (conv4), while FPN [26] uses a fully connected head (2-fc) on multiple level feature maps. However, there is a lack of **understanding** between the two head structures with respect to the two tasks (object classification and localization).

In this paper, we perform a thorough comparison between the fully connected head (fc-head) and the convolution head (conv-head) on the two detection tasks, i.e. object classification and localization. We find that *these two different head structures are complementary*. fc-head is more

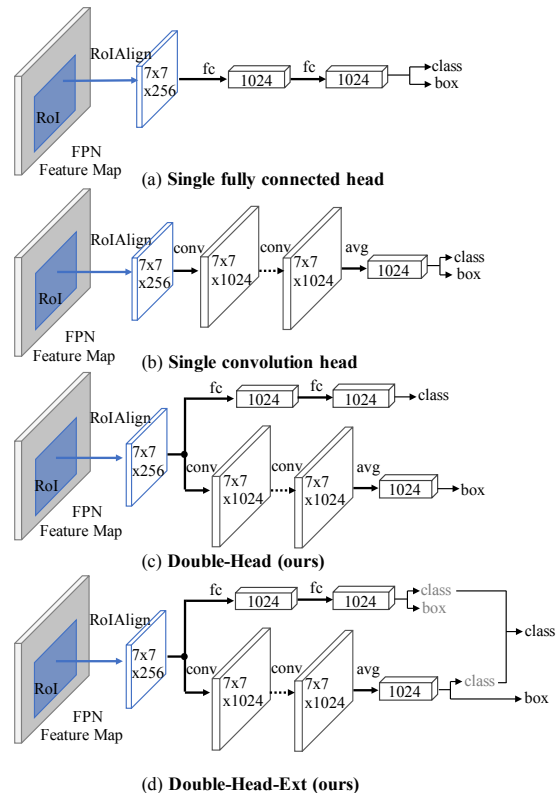


Figure 1. Comparison between single head and double heads, (a) a single fully connected (2-fc) head, (b) a single convolution head, (c) Double-Head, which splits classification and localization on a fully connected head and a convolution head respectively, and (d) Double-Head-Ext, which extends Double-Head by introducing supervision from unfocused tasks during training and combining classification scores from both heads during inference.

suitable for the classification task as its classification score is more correlated to the intersection over union (IoU) between a proposal and its corresponding ground truth box. Meanwhile, *conv-head* provides more accurate bounding box regression.

We believe this is because *fc-head* is spatially sensitive, having different parameters for different parts of a proposal, while *conv-head* shares convolution kernels for all parts. To validate this, we examine the output feature maps of both

heads and confirm that *fc-head* is more spatially sensitive. As a result, *fc-head* is better to distinguish between a complete object and part of an object (classification) and *conv-head* is more robust to regress the whole object (bounding box regression).

In light of above findings, we propose a Double-Head method, which includes a fully connected head (*fc-head*) for classification and a convolution head (*conv-head*) for bounding box regression (see Figure 1-(c)), to leverage advantages of both heads. This design outperforms both single *fc-head* and single *conv-head* (see Figure 1-(a), (b)) by a non-negligible margin. In addition, we extend Double-Head (Figure 1-(d)) to further improve the accuracy by leveraging unfocused tasks (i.e. classification in *conv-head*, and bounding box regression in *fc-head*). Our method outperforms FPN baseline by a non-negligible margin on MS COCO 2017 dataset [28], gaining 3.5 and 2.8 AP for using ResNet-50 and ResNet-101 backbones, respectively.

## 2. Related Work

**One-stage Object Detectors:** OverFeat [37] detects objects by sliding windows on feature maps. SSD [29, 9] and YOLO [32, 33, 34] have been tuned for speed by predicting object classes and locations directly. RetinaNet [27] alleviates the extreme foreground-background class imbalance problem by introducing focal loss. Point-based methods [21, 22, 47, 7, 48] model an object as keypoints (corner, center, etc), and are built on keypoint estimation networks.

**Two-stage Object Detectors:** RCNN [12] applies a deep neural network to extract features from proposals generated by selective search [42]. SPPNet [14] speeds up RCNN significantly using spatial pyramid pooling. Fast RCNN [10] improves the speed and performance utilizing a differentiable RoI Pooling. Faster RCNN [35] introduces Region Proposal Network (RPN) to generate proposals. R-FCN [4] employs position sensitive RoI pooling to address the translation-variance problem. FPN [26] builds a top-down architecture with lateral connections to extract features across multiple layers.

**Backbone Networks:** Fast RCNN [10] and Faster RCNN [35] extract features from conv4 of VGG-16 [38], while FPN [26] utilizes features from multiple layers (conv2 to conv5) of ResNet [15]. Deformable ConvNets [5, 49] propose deformable convolution and deformable Region of Interests (RoI) pooling to augment spatial sampling locations. Trident Network [24] generates scale-aware feature maps with multi-branch architecture. MobileNet [17, 36] and ShuffleNet [46, 30] introduce efficient operators (like depth-wise convolution, group convolution, channel shuffle, etc) to speed up on mobile devices.

**Detection Heads:** Light-Head RCNN [25] introduces an efficient head network with thin feature maps. Cascade RCNN [3] constructs a sequence of detection heads trained

with increasing IoU thresholds. Feature Sharing Cascade RCNN [23] utilizes feature sharing to ensemble multi-stage outputs from Cascade RCNN [3] to improve the results. Mask RCNN [13] introduces an extra head for instance segmentation. COCO Detection 18 Challenge winner (Megvii) [1] couples bounding box regression and instance segmentation in a convolution head. IoU-Net [20] introduces a branch to predict IoUs between detected bounding boxes and their corresponding ground truth boxes. Similar to IoU-Net, Mask Scoring RCNN [18] presents an extra head to predict Mask IoU scores for each segmentation mask. He et. al. [16] learns uncertainties of bounding box prediction with an extra task to improve the localization results. Learning-to-Rank [39] utilizes an extra head to produce a rank value of a proposal for Non-Maximum Suppression (NMS). Zhang and Wang [45] point out that there exist misalignments between classification and localization task domains. In contrast to existing methods, which apply a single head to extract Region of Interests (RoI) features for both classification and bounding box regression tasks, we propose to split these two tasks into different heads, based upon our thorough analysis.

## 3. Analysis: Comparison between *fc-head* and *conv-head*

In this section, we compare *fc-head* and *conv-head* for both classification and bounding box regression. For each head, we train a model with FPN backbone [26] using ResNet-50 [15] on MS COCO 2017 dataset [28]. The *fc-head* includes two fully connected layers. The *conv-head* has five residual blocks. The evaluation and analysis is conduct on the MS COCO 2017 validation set with 5,000 images. *fc-head* and *conv-head* have 36.8% and 35.9% AP, respectively.

### 3.1. Data Processing for Analysis

To make a fair comparison, we perform analysis for both heads on predefined proposals rather than proposals generated by RPN [35], as the two detectors have different proposals. The predefined proposals include sliding windows around the ground truth box with different sizes. For each ground truth object, we generate about 14,000 proposals. The IoUs between these proposals and the ground truth box (*denoted as proposal IoUs*) gradually change from zero (background) to one (the ground truth box). For each proposal, both detectors (*fc-head* and *conv-head*) generate classification scores and regressed bounding boxes. This process is applied for all objects in the validation set.

We split the IoUs between predefined proposals and their corresponding ground truth into 20 bins uniformly, and group these proposals accordingly. For each group, we calculate mean and standard deviation of classification scores

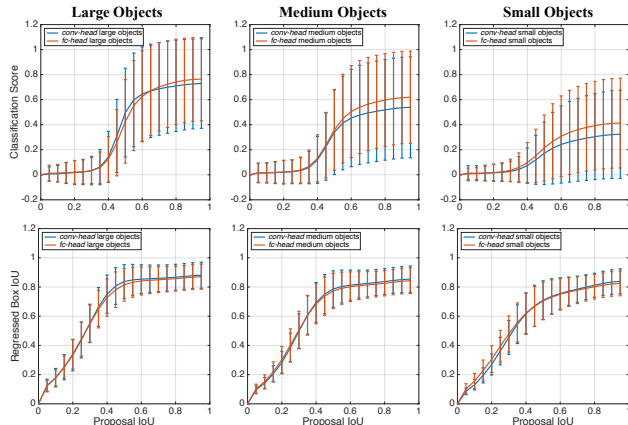


Figure 2. Comparison between *fc-head* and *conv-head*. Top row: mean and standard deviation of classification scores. Bottom row: mean and standard deviation of IoUs between regressed boxes and their corresponding ground truth. Classification scores in *fc-head* are more correlated to proposal IoUs than in *conv-head*. *conv-head* has better regression results than *fc-head*.

and IoUs of regressed boxes. Figure 2 shows the results for small, medium and large objects.

### 3.2. Comparison on Classification Task

The first row of Figure 2 shows the classification scores for both *fc-head* and *conv-head*. Compared to *conv-head*, *fc-head* provides higher scores for proposals with higher IoUs. This indicates that *classification scores of fc-head are more correlated to IoUs between proposals and corresponding ground truth than of conv-head*, especially for small objects. To validate this, we compute the Pearson correlation coefficient (PCC) between proposal IoUs and classification scores. The results (shown in Figure 3 (Left)) demonstrate that the classification scores of *fc-head* are more correlated to the proposal IoUs.

We also compute Pearson correlation coefficient for the proposals generated by RPN [35] and final detected boxes after NMS. Results are shown in Figure 3 (Right). Similar to the predefined proposals, *fc-head* has higher PCC than *conv-head*. Thus, the detected boxes with higher IoUs are ranked higher when calculating AP due to their higher classification scores.

### 3.3. Comparison on Localization Task

The second row of Figure 2 shows IoUs between the regressed boxes and their corresponding ground truth for both *fc-head* and *conv-head*. Compared to *fc-head*, the regressed boxes of *conv-head* are more accurate when the proposal IoU is above 0.4. This demonstrates that *conv-head has better regression ability than fc-head*.

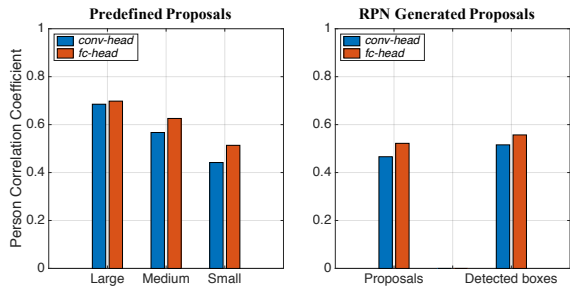


Figure 3. Pearson correlation coefficient (PCC) between classification scores and IoUs. Left: PCC of predefined proposals for large, medium and small objects. Right: PCC of proposals generated by RPN and detected boxes after NMS.

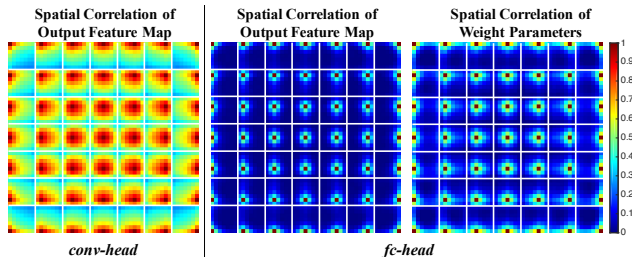


Figure 4. Left: Spatial correlation in output feature map of *conv-head*. Middle: Spatial correlation in output feature map of *fc-head*. Right: Spatial correlation in weight parameters of *fc-head*. *conv-head* has significantly more spatial correlation in output feature map than *fc-head*. *fc-head* has a similar spatial correlation pattern in output feature map and weight parameters.

### 3.4. Discussion

*Why does fc-head show more correlation between the classification scores and proposal IoUs, and perform worse in localization?* We believe it is because *fc-head* is more spatially sensitive than *conv-head*. Intuitively, *fc-head* applies *unshared* transformations (fully connected layer) over different positions of the input feature map. Thus, the spatial information is implicitly embedded. The spatial sensitivity of *fc-head* helps distinguish between a complete object and part of an object, but is not robust to determine the offset of the whole object. In contrast, *conv-head* uses a *shared* transformation (convolutional kernels) on all positions of the input feature map, and uses average pooling to aggregate.

Next, we inspect the spatial sensitivity of *conv-head* and *fc-head*. For *conv-head* whose output feature map is a  $7 \times 7$  grid, we compute the spatial correlation between any pair of locations using the cosine distance between the corresponding two feature vectors. This results in a  $7 \times 7$  correlation matrix per cell, representing the correlation between the current cell and other cells. Thus, the spatial correlation of a output feature map can be visualized by tiling the correlation matrices of all cells in a  $7 \times 7$  grid. Figure 4 (Left)

shows the average spatial correlation of *conv-head* over multiple objects. For *fc-head* whose output is not a feature map, but a feature vector with dimension 1024, we reconstruct its output feature map. This can be done by splitting the weight matrix of fully connected layer ( $256 \cdot 7 \cdot 7 \times 1024$ ) by spatial locations. Each cell in the  $7 \times 7$  grid has a transformation matrix with dimension  $256 \times 1024$ , which is used to generate output features for that cell. Thus, output feature map  $7 \times 7 \times 1024$  for *fc-head* is reconstructed. Then we can compute its spatial correlation in a similar manner to *conv-head*. Figure 4 (Middle) shows the average spatial correlation in output feature map of *fc-head* over multiple objects. *fc-head* has significant less spatial correlation than *conv-head*. This supports our conjecture that *fc-head* is more spatially sensitive than *conv-head*, making it easier to distinguish if one proposal covers one complete or partial object. On the other hand, it is not as robust as *conv-head* to regress bounding boxes.

We further examine the spatial correlation of weight parameters ( $256 \cdot 7 \cdot 7 \times 1024$ ) in *fc-head*, by splitting them along spatial locations. As a result, each cell of the  $7 \times 7$  grid has a matrix with dimension  $256 \times 1024$ , which is used to compute correlation with other cells. Similar to the correlation analysis on output feature map, we compute the correlation matrices for all cells. Figure 4 (Right) shows the spatial correlation in weight parameters of *fc-head*. It has a similar pattern to the spatial correlation in output feature map of *fc-head* (shown in Figure 4 (Middle)).

## 4. Our Approach: Double-Head

Based upon above analysis, we propose a double-head method to leverage the advantages of two head structures. In this section, we firstly introduce the network structure of Double-Head, which has a fully connected head (*fc-head*) for classification and a convolution head (*conv-head*) for bounding box regression. Then, we extend Double-Head to Double-Head-Ext by leveraging unfocused tasks (i.e. bounding box regression in *fc-head* and classification in *conv-head*).

### 4.1. Network Structure

Our Double-Head method (see Figure 1-(c)) splits classification and localization into *fc-head* and *conv-head*, respectively. The details of backbone and head networks are described as follows:

**Backbone:** We use FPN [26] backbone to generate region proposals and extract object features from multiple levels using RoIAlign [13]. Each proposal has a feature map with size  $256 \times 7 \times 7$ , which is transformed by *fc-head* and *conv-head* into two feature vectors (each with dimension 1024) for classification and bounding box regression, respectively.

**Fully Connected Head (*fc-head*)** has two fully connected layers (see Figure 1-(c)), following the design in FPN [26]

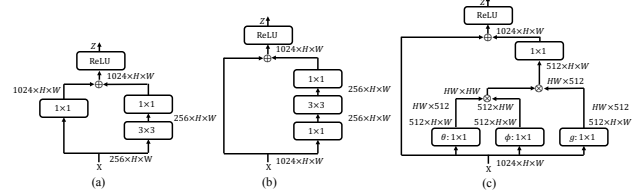


Figure 5. Network architectures of three components: (a) residual block to increase the number of channels (from 256 to 1024), (b) residual bottleneck block, and (c) non-local block.

(Figure 1-(a)). The output dimension is 1024. The parameter size is 13.25M.

**Convolution Head (*conv-head*)** stacks  $K$  residual blocks [15]. The first block increases the number of channels from 256 to 1024 (shown in Figure 5-(a)), and others are bottleneck blocks [15] (shown in Figure 5-(b)). At the end, average pooling is used to generate the feature vector with dimension 1024. Each residual block has 1.06M parameters. We also introduce a variation for the convolution head by inserting a non-local block [43] (see Figure 5-(c)) before each bottleneck block to enhance foreground objects. Each non-local block has 2M parameters.

**Loss Function:** Both heads (i.e. *fc-head* and *conv-head*) are jointly trained with region proposal network (RPN) end to end. The overall loss is computed as follows:

$$\mathcal{L} = \omega^{fc} \mathcal{L}^{fc} + \omega^{conv} \mathcal{L}^{conv} + \mathcal{L}^{rpn}, \quad (1)$$

where  $\omega^{fc}$  and  $\omega^{conv}$  are weights for *fc-head* and *conv-head*, respectively.  $\mathcal{L}^{fc}$ ,  $\mathcal{L}^{conv}$ ,  $\mathcal{L}^{rpn}$  are the losses for *fc-head*, *conv-head* and RPN, respectively.

### 4.2. Extension: Leveraging Unfocused Tasks

In vanilla Double-Head, each head focuses on its assigned task (i.e. classification in *fc-head* and bounding box regression in *conv-head*). In addition, we found that unfocused tasks (i.e. bounding box regression in *fc-head* and classification in *conv-head*) are helpful in two aspects: (a) bounding box regression provides auxiliary supervision for *fc-head*, and (b) classifiers from both heads are complementary. Therefore, we introduce unfocused task supervision in training and propose a complementary fusion method to combine classification scores from both heads during inference (see Figure 1-(d)). This extension is referred to as Double-Head-Ext.

**Unfocused Task Supervision:** Due to the introduction of the unfocused tasks, the loss for *fc-head* ( $\mathcal{L}^{fc}$ ) includes both classification loss and bounding box regression loss as follows:

$$\mathcal{L}^{fc} = \lambda^{fc} L_{cls}^{fc} + (1 - \lambda^{fc}) L_{reg}^{fc}, \quad (2)$$

where  $L_{cls}^{fc}$  and  $L_{reg}^{fc}$  are the classification and bounding box regression losses in *fc-head*, respectively.  $\lambda^{fc}$  is the weight

that controls the balance between the two losses in *fc-head*. In the similar manner, we define the loss for the convolution head ( $\mathcal{L}^{conv}$ ) as follows:

$$\mathcal{L}^{conv} = (1 - \lambda^{conv})L_{cls}^{conv} + \lambda^{conv}L_{reg}^{conv}, \quad (3)$$

where  $L_{cls}^{conv}$  and  $L_{reg}^{conv}$  are classification and bounding box regression losses in *conv-head*, respectively. Different from  $\lambda^{fc}$  that is multiplied by the classification loss  $L_{cls}^{fc}$ , the balance weight  $\lambda^{conv}$  is multiplied by the regression loss  $L_{reg}^{conv}$ , as the bounding box regression is the focused task in *conv-head*. Note that the vanilla Double-Head is a special case when  $\lambda^{fc} = 1$  and  $\lambda^{conv} = 1$ . Similar to FPN [26], cross entropy loss is applied to classification, and Smooth- $L_1$  loss is used for bounding box regression.

**Complementary Fusion of Classifiers:** We believe that the two heads (i.e. *fc-head* and *conv-head*) capture complementary information for object classification due to their different structures. Therefore we propose to fuse the two classifiers as follows:

$$s = s^{fc} + s^{conv}(1 - s^{fc}) = s^{conv} + s^{fc}(1 - s^{conv}), \quad (4)$$

where  $s^{fc}$  and  $s^{conv}$  are classification scores from *fc-head* and *conv-head*, respectively. The increment from the first score (e.g.  $s^{fc}$ ) is a product of the second score and the reverse of the first score (e.g.  $s^{conv}(1 - s^{fc})$ ). This is different from [3] which combining all classifiers by average. Note that this fusion is only applicable when  $\lambda^{fc} \neq 0$  and  $\lambda^{conv} \neq 1$ .

## 5. Experimental Results

We evaluate our approach on MS COCO 2017 dataset [28] and Pascal VOC07 dataset [8]. MS COCO 2017 dataset has 80 object categories. We train on `train2017` (118K images) and report results on `val2017` (5K images) and `test-dev` (41K images). The standard COCO-style Average Precision (AP) with different IoU thresholds from 0.5 to 0.95 is used as evaluation metric. Pascal VOC07 dataset has 20 object categories. We train on `trainval` with 5K images and report results on `test` with 5K images. We perform ablation studies to analyze different components of our approach, and compare our approach to base-lines and state-of-the-art.

### 5.1. Implementation Details

Our implementation is based on Mask R-CNN benchmark in Pytorch 1.0 [31]. Images are resized such that the shortest side is 800 pixels. We use no data augmentation for testing, and only horizontal flipping augmentation for training. The implementation details are described as follows:

**Architecture:** Our approach is evaluated on two FPN [26] backbones (ResNet-50 and ResNet-101 [15]), which are pretrained on ImageNet [6]. The standard RoI pooling is

replaced by RoIAlign [13]. Both heads and RPN are jointly trained end to end. All batch normalization (BN) [19] layers in the backbone are frozen. Each convolution layer in *conv-head* is followed by a BN layer. The bounding box regression is class-specific.

**Hyper-parameters:** All models are trained using 4 NVIDIA P100 GPUs with 16GB memory, and a mini-batch size of 2 images per GPU. The weight decay is 1e-4 and momentum is 0.9.

**Learning Rate Scheduling:** All models are fine-tuned with 180k iterations. The learning rate is initialized with 0.01 and reduced by 10 after 120K and 160K iterations, respectively.

### 5.2. Ablation Study

We perform a number of ablations to analyze Double-Head with ResNet-50 backbone on COCO `val2017`.

**Double-Head Variations:** Four variations of double heads are compared:

- **Double-FC** splits the classification and bounding box regression into two fully connected heads, which have the identical structure.
- **Double-Conv** splits the classification and bounding box regression into two convolution heads, which have the identical structure.
- **Double-Head** includes a fully connected head (*fc-head*) for classification and a convolution head (*conv-head*) for bounding box regression.
- **Double-Head-Reverse** switches tasks between two heads (i.e. *fc-head* for bounding box regression and *conv-head* for classification), compared to Double-Head.

The detection performances are shown in Table 1. The top group shows performances for single head detectors. The middle group shows performances for detectors with double heads. The weight for each loss (classification and bounding box regression) is set to 1.0. Compared to the middle group, the bottom group uses different loss weights for *fc-head* and *conv-head* ( $\omega^{fc} = 2.0$  and  $\omega^{conv} = 2.5$ ), which are set empirically.

Double-Head outperforms single head detectors by a non-negligible margin (2.0+ AP). It also outperforms Double-FC and Double-Conv by at least 1.4 AP. Double-Head-Reverse has the worst performance (drops 6.2+ AP compared to Double-Head). This validates our findings that *fc-head* is more suitable for classification, while *conv-head* is more suitable for localization.

Single-Conv performs better than Double-Conv. We believe that the regression task helps the classification task when sharing a single convolution head. This is supported by sliding window analysis (see details in section 3.1).

	<i>fc-head</i>		<i>conv-head</i>		AP
	cls	reg	cls	reg	
Single-FC	1.0	1.0	-	-	36.8
Single-Conv	-	-	1.0	1.0	35.9
Double-FC	1.0	1.0	-	-	37.3
Double-Conv	-	-	1.0	1.0	33.8
Double-Head-Reverse	-	1.0	1.0	-	32.6
Double-Head	1.0	-	-	1.0	<b>38.8</b>
Double-FC	2.0	2.0	-	-	38.1
Double-Conv	-	-	2.5	2.5	34.3
Double-Head-Reverse	-	2.0	2.5	-	32.0
Double-Head	2.0	-	-	2.5	<b>39.5</b>

Table 1. Evaluations of detectors with different head structures on COCO val2017. The backbone is FPN with ResNet-50. The top group shows performances for single head detectors. The middle group shows performances for detectors with double heads. The weight for each loss (classification and bounding box regression) is set to 1.0. Compared to the middle group, the bottom group uses different loss weight for *fc-head* and *conv-head* ( $\omega^{fc} = 2.0$ ,  $\omega^{conv} = 2.5$ ). Clearly, Double-Head has the best performance, outperforming others by a non-negligible margin. Double-Head-Reverse has the worst performance.

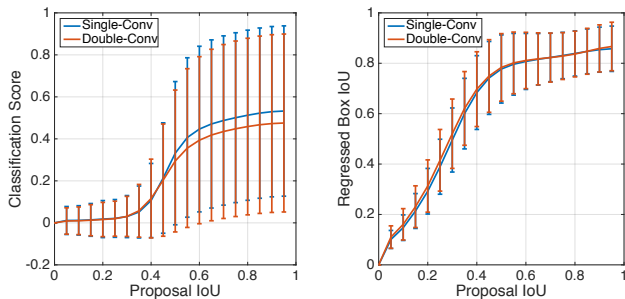


Figure 6. Comparison between Single-Conv and Double-Conv. Left: mean and standard deviation of classification scores. Right: mean and standard deviation of IoUs between regressed boxes and their corresponding ground truth. Single-Conv has higher classification scores than Double-Conv, while regression results are comparable.

Figure 6 shows the comparison between Single-Conv and Double-Conv. Their regression results are comparable. But Single-Conv has higher classification scores than Double-Conv on proposals which have higher IoUs with the ground truth box. Thus, sharing regression and classification on a single convolution head encourages the correlation between classification scores and proposal IoUs. This allows Single-Conv to better determine if a complete object is covered.

In contrast, sharing two tasks in a single fully connected head (Single-FC) is not as good as separating them in two heads (Double-FC). We believe that adding the regression task in the same head with equal weight introduces confliction. This is supported by the sliding window analysis. Figure 7 shows that Double-FC has slightly higher classification scores and higher IoUs between the regressed boxes

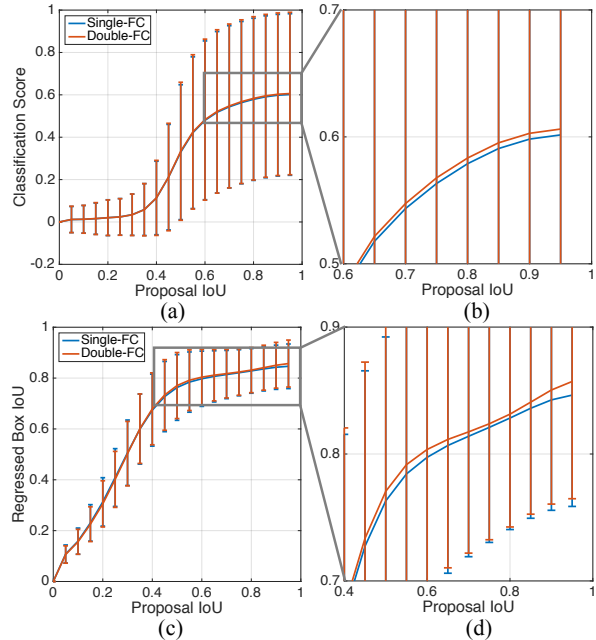


Figure 7. Comparison between Single-FC and Double-FC. (a): mean and standard deviation of classification scores. (b): zooming in of the box in plot-(a). (c): mean and standard deviation of IoUs between regressed boxes and their corresponding ground truth. (d): zooming in of the box in plot-(c). Double-FC has slightly higher classification scores and better regression results than Single-FC.

and their corresponding ground truth than Single-FC.

**Depth of *conv-head*:** We study the number of blocks for the convolution head. The evaluations are shown in Table 2. The first group has  $K$  residual blocks (Figure 5-(a-b)), while the second group has alternating  $(K + 1)/2$  residual blocks and  $(K - 1)/2$  non-local blocks (Figure 5-(c)). When using a single block in *conv-head*, the performance is slightly behind FPN baseline (drops 0.1 AP) as it is too shallow. However, adding another convolution block boosts the performance substantially (gains 1.9 AP from FPN baseline). As the number of blocks increases, the performance improves gradually with decreasing growth rate. Considering the trade-off between accuracy and complexity, we choose *conv-head* with 3 residual blocks and 2 non-local blocks ( $K = 5$  in the second group) for the rest of the paper, which gains 3.0 AP from baseline.

**More training iterations:** When increasing training iterations from 180k ( $1\times$  training) to 360k ( $2\times$  training), Double-Head gains 0.6 AP (from 39.8 to 40.4).

**Balance Weights  $\lambda^{fc}$  and  $\lambda^{conv}$ :** Figure 8 shows APs for different choices of  $\lambda^{fc}$  and  $\lambda^{conv}$ . For each  $(\lambda^{fc}, \lambda^{conv})$  pair, we train a Double-Head-Ext model. The vanilla Double-Head model is corresponding to  $\lambda^{fc} = 1$  and  $\lambda^{conv} = 1$ , while other models involve supervision from

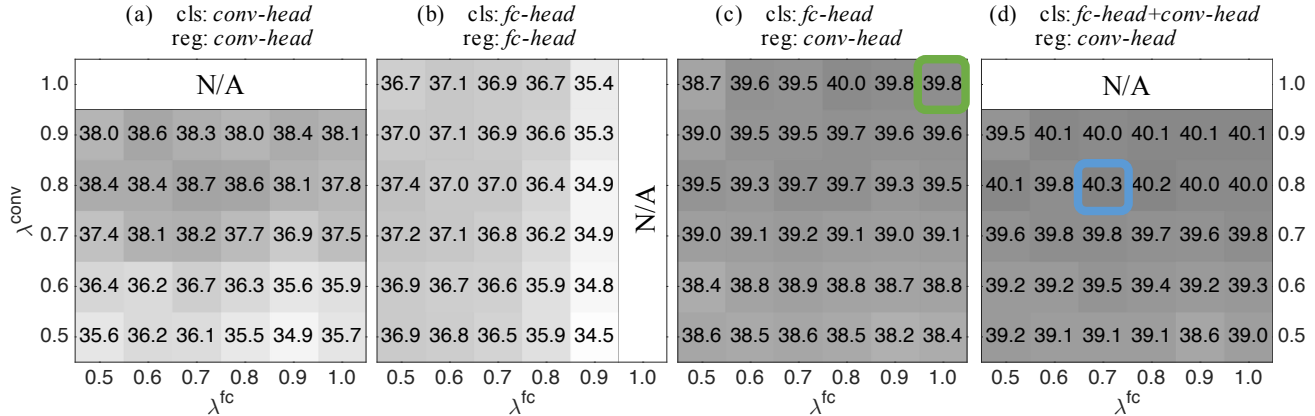


Figure 8. AP over balance weights  $\lambda^{fc}$  and  $\lambda^{conv}$ . For each  $(\lambda^{fc}, \lambda^{conv})$  pair, we trained a Double-Head-Ext model. Note that the vanilla Double-Head is a special case with  $\lambda^{fc} = 1, \lambda^{conv} = 1$ . For each model, we evaluate AP in four ways: (a) using *conv-head* alone, (b) using *fc-head* alone, (c) using classification from *fc-head* and bounding box from *conv-head*, and (d) using classification fusion from both heads and bounding box from *conv-head*. Note that the first row in (a) and (d) is not available, due to the unavailability of classification in *conv-head* when  $\lambda^{conv} = 1$ . The last column in (b) is not available, due to the unavailability of bound box regression in *fc-head* when  $\lambda^{fc} = 1$ .

NL	$K$	param	AP	AP <sub>0.5</sub>	AP <sub>0.75</sub>
	0	-	36.8	58.7	40.4
	1	1.06M	36.7 (-0.1)	59.3	39.6
	2	2.13M	38.7 (+1.9)	59.2	41.9
	3	3.19M	39.2 (+2.4)	59.4	42.5
	4	4.25M	39.3 (+2.5)	59.2	42.9
	5	5.31M	39.5 (+2.7)	59.6	43.2
	6	6.38M	39.5 (+2.7)	59.4	43.3
	7	7.44M	39.7 (+2.9)	59.8	43.2
✓	3	4.13M	38.8 (+2.0)	59.2	42.4
✓	5	7.19M	39.8 (+3.0)	59.6	43.6
✓	7	10.25M	<b>40.0 (+3.2)</b>	<b>59.9</b>	<b>43.7</b>

Table 2. The number of blocks (Figure 5) in the convolution head. The baseline ( $K = 0$ ) is equivalent to the original FPN [26] which uses *fc-head* alone. The first group only stacks residual blocks, while the second group alternates  $(K + 1)/2$  residual blocks and  $(K - 1)/2$  non-local blocks.

Fusion Method	AP	AP <sub>0.5</sub>	AP <sub>0.75</sub>
No fusion	39.7	59.5	43.4
Max	39.9	59.7	43.7
Average	40.1	59.8	44.1
Complementary	<b>40.3</b>	<b>60.3</b>	<b>44.2</b>

Table 3. Fusion of classifiers from both heads. Complementary fusion (Eq. 4) outperforms others. The model is trained using weights  $\lambda^{fc} = 0.7, \lambda^{conv} = 0.8$ .

unfocused tasks. For each model, we evaluate AP for using *conv-head* alone (Figure 8-(a)), using *fc-head* alone (Figure 8-(b)), using classification from *fc-head* and bounding box from *conv-head* (Figure 8-(c)), and using classification fusion from both heads and bounding box from *conv-head* (Figure 8-(d)).  $\omega^{fc}$  and  $\omega^{conv}$  are set as 2.0 and 2.5 in all

Method	AP	AP <sub>0.5</sub>	AP <sub>0.75</sub>
FPN baseline [26]	47.4	75.7	41.9
Double-Head-Ext (ours)	<b>49.2</b>	<b>76.7</b>	<b>45.6</b>

Table 4. Comparisons with FPN baseline [26] on VOC07 datasets with ResNet-50 backbone. Our Double-Head-Ext outperforms FPN baseline.

experiments, respectively.

We summarize key observations as follows. Firstly, using two heads (Figure 8-(c)) outperforms using a single head (Figure 8-(a), (b)) for all  $(\lambda^{fc}, \lambda^{conv})$  pairs by at least 0.9 AP. Secondly, fusion of classifiers introduces at least additional 0.4 AP improvement for all  $(\lambda^{fc}, \lambda^{conv})$  pairs (compare Figure 8-(c) and (d)). And finally the unfocused tasks are helpful as the best Double-Head-Ext model (40.3 AP) is corresponding to  $\lambda^{fc} = 0.7, \lambda^{conv} = 0.8$  (blue box in Figure 8-(d)). It outperforms Double-Head (39.8 AP, green box in Figure 8-(c)) without using unfocused tasks by 0.5 AP. For the rest of the paper, we use  $\lambda^{fc} = 0.7$  and  $\lambda^{conv} = 0.8$  for Double-Head-Ext.

**Fusion of Classifiers:** We study three different ways to fuse the classification scores from both the fully connected head ( $s^{fc}$ ) and the convolution head ( $s^{conv}$ ) during inference: (a) average, (b) maximum, and (c) complementary fusion using Eq. (4). The evaluations are shown in Table 3. The proposed complementary fusion outperforms other fusion methods (max and average) and gains 0.6 AP compared to using the score from *fc-head* alone.

Method	Backbone	AP	AP <sub>0.5</sub>	AP <sub>0.75</sub>	AP <sub>s</sub>	AP <sub>m</sub>	AP <sub>l</sub>
Faster R-CNN [35]	ResNet-50-C4	34.8	55.8	37.0	19.1	38.8	48.2
FPN baseline [26]	ResNet-50	36.8	58.7	40.4	21.2	40.1	48.8
Double-Head (ours)	ResNet-50	39.8	59.6	43.6	<b>22.7</b>	42.9	53.1
Double-Head-Ext (ours)	ResNet-50	<b>40.3</b>	<b>60.3</b>	<b>44.2</b>	22.4	<b>43.3</b>	<b>54.3</b>
Faster R-CNN [35]	ResNet-101-C4	38.5	59.4	41.4	19.7	43.1	53.3
FPN baseline [26]	ResNet-101	39.1	61.0	42.4	22.2	42.5	51.0
Double-Head (ours)	ResNet-101	41.5	61.7	45.6	23.8	<b>45.2</b>	54.9
Double-Head-Ext (ours)	ResNet-101	<b>41.9</b>	<b>62.4</b>	<b>45.9</b>	<b>23.9</b>	<b>45.2</b>	<b>55.8</b>

Table 5. Object detection results (bounding box AP) on COCO val2017. Note that FPN baseline only has *fc-head*. Our Double-Head and Double-Head-Ext outperform both Faster R-CNN and FPN baselines on two backbones (ResNet-50 and ResNet-101).

Method	Backbone	AP	AP <sub>0.5</sub>	AP <sub>0.75</sub>	AP <sub>s</sub>	AP <sub>m</sub>	AP <sub>l</sub>
FPN [26]	ResNet-101	36.2	59.1	39.0	18.2	39.0	48.2
Mask RCNN [13]	ResNet-101	38.2	60.3	41.7	20.1	41.1	50.2
Deep Regionlets [44]	ResNet-101	39.3	59.8	-	21.7	43.7	50.9
IOU-Net [20]	ResNet-101	40.6	59.0	-	-	-	-
Soft-NMS [2]	Aligned-Inception-ResNet	40.9	<b>62.8</b>	-	23.3	43.6	53.3
LTR [39]	ResNet-101	41.0	60.8	44.5	23.2	44.5	52.5
Fitness NMS [41]	DeNet-101 [40]	41.8	60.9	44.9	21.5	<b>45.0</b>	<b>57.5</b>
Double-Head-Ext (ours)	ResNet-101	<b>42.3</b>	<b>62.8</b>	<b>46.3</b>	<b>23.9</b>	44.9	54.3

Table 6. Object detection results (bounding box AP), vs. state-of-the-art on COCO test-dev. All methods are in the family of two-stage detectors with a single training stage. Our Double-Head-Ext achieves the best performance.

### 5.3. Main Results

**Comparison with Baselines on VOC07:** We conduct experiments on Pascal VOC07 dataset and results are shown in Table 4. Compared with FPN, our method gains 1.8 AP. Specifically, it gains 3.7 AP<sub>0.75</sub> on the higher IoU threshold 0.75 and gains 1.0 AP<sub>0.5</sub> on the lower IoU threshold 0.5.

**Comparison with Baselines on COCO:** Table 5 shows the comparison between our method with Faster RCNN [35] and FPN [26] baselines on COCO val2017. Our method outperforms both baselines on *all* evaluation metrics. Compared with FPN, our Double-Head-Ext gains 3.5 and 2.8 AP on ResNet-50 and ResNet-101 backbones, respectively. Specifically, our method gains 3.5+ AP on the higher IoU threshold (0.75) and 1.4+ AP on the lower IoU threshold (0.5) for both backbones. This demonstrates the advantage of our method with double heads.

We also observe that Faster R-CNN and FPN have different preferences over object sizes when using ResNet-101 backbone: i.e. Faster R-CNN has better AP on medium and large objects, while FPN is better on small objects. Even comparing with the best performance among FPN and Faster R-CNN across different sizes, our Double-Head-Ext gains 1.7 AP on small objects, 2.1 AP on medium objects and 2.5 AP on large objects. This demonstrates the superiority of our method, which leverages the advantage of *fc-head* on classification and the advantage of *conv-head* on localization.

**Comparison with State-of-the-art on COCO:** We compare our Double-Head-Ext with the state-of-the-art methods on MS COCO 2017 test-dev in Table 6. ResNet-101 is

used as the backbone. For fair comparison, the performance of single-model inference is reported for all methods. Here, we only consider the two-stage detectors with a single training stage. Our Double-Head-Ext achieves the best performance with 42.3 AP. This demonstrates the superior performance of our method. Note that Cascade RCNN [3] is not included as it involves multiple training stages. Even through our method only has one training stage, the performance of our method is slightly below Cascade RCNN (42.8 AP).

## 6. Conclusions

In this paper, we perform a thorough analysis and find an interesting fact that two widely used head structures (convolution head and fully connected head) have opposite preferences towards classification and localization tasks in object detection. Specifically, *fc-head* is more suitable for the classification task, while *conv-head* is more suitable for the localization task. Furthermore, we examine the output feature maps of both heads and find that *fc-head* has more spatial sensitivity than *conv-head*. Thus, *fc-head* has more capability to distinguish a complete object from part of an object, but is not robust to regress the whole object. Based upon these findings, we propose a Double-Head method, which has a fully connected head focusing on classification and a convolution head for bounding box regression. Without bells and whistles, our method gains +3.5 and +2.8 AP on MS COCO dataset from FPN baselines with ResNet-50 and ResNet-101 backbones, respectively. We hope that our findings are helpful for future research in object detection.



## References

- [1] Mscoco instance segmentation challenges 2018 megvii (face++) team. <http://presentations.cocodataset.org/ECCV18/COCO18-Detect-Megvii.pdf>, 2018.
- [2] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms—improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017.
- [3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [4] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems 29*, pages 379–387, 2016.
- [5] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 764–773, 2017.
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [7] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
- [9] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Amrbrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.
- [10] Ross Girshick. Fast R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015.
- [11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.
- [12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.
- [13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European conference on computer vision*, pages 346–361. Springer, 2014.
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [16] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2888–2897, 2019.
- [17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [18] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask Scoring R-CNN. In *CVPR*, 2019.
- [19] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.
- [20] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yun-jing Jiang. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 784–799, 2018.
- [21] Hei Law and Jia Deng. CornerNet: Detecting Objects as Paired Keypoints. In *ECCV*, 2018.
- [22] Hei Law, Yun Teng, Olga Russakovsky, and Jia Deng. Cornernet-lite: Efficient keypoint based object detection. *arXiv preprint arXiv:1904.08900*, 2019.
- [23] Ang Li, Xue Yang, and Chongyang Zhang. Rethinking classification and localization for cascade r-cnn. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019.
- [24] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [25] Zeming Li, Chao Peng, Gang Yu, Xiangyu Zhang, Yangdong Deng, and Jian Sun. Light-head r-cnn: In defense of two-stage object detector. *arXiv preprint arXiv:1711.07264*, 2017.
- [26] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, July 2017.
- [27] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

- [30] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 116–131, 2018.
- [31] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. <https://github.com/facebookresearch/maskrcnn-benchmark>, 2018.
- [32] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [33] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.
- [34] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.
- [35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [36] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.
- [37] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Robert Fergus, and Yann Lecun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations (ICLR2014), CBLS, April 2014*, 2014.
- [38] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.
- [39] Zhiyu Tan, Xuecheng Nie, Qi Qian, Nan Li, and Hao Li. Learning to rank proposals for object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [40] Lachlan Tychsen-Smith and Lars Petersson. Denet: Scalable real-time object detection with directed sparse sampling. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 428–436, 2017.
- [41] Lachlan Tychsen-Smith and Lars Petersson. Improving object localization with fitness nms and bounded iou loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6877–6885, 2018.
- [42] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.
- [43] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.
- [44] Hongyu Xu, Xutao Lv, Xiaoyu Wang, Zhou Ren, Navaneeth Bodla, and Rama Chellappa. Deep regionlets for object detection. In *The European Conference on Computer Vision (ECCV)*, September 2018.
- [45] Haichao Zhang and Jianyu Wang. Towards adversarially robust object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [46] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 2018.
- [47] Xingyi Zhou, Dequan Wang, and Philipp Krahenbuhl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.
- [48] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krähenbühl. Bottom-up object detection by grouping extreme and center points. In *CVPR*, 2019.
- [49] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. *CVPR*, 2019.