

CenterNet: Keypoint Triplets for Object Detection

Kaiwen Duan^{1*} Song Bai² Lingxi Xie³ Honggang Qi¹ Qingming Huang¹ Qi Tian³

¹University of Chinese Academy of Sciences ²University of Oxford ³Huawei Noah’s Ark Lab

kaiwen.duan@vipl.ict.ac.cn songbai.site@gmail.com 198808xc@gmail.com hgqi@ucas.ac.cn
qmhuang@ucas.ac.cn tian.qi1@huawei.com

Abstract

In object detection, keypoint-based approaches often suffer a large number of incorrect object bounding boxes, arguably due to the lack of an additional look into the cropped regions. This paper presents an efficient solution which explores the visual patterns within each cropped region with minimal costs. We build our framework upon a representative one-stage keypoint-based detector named CornerNet. Our approach, named CenterNet, detects each object as a triplet, rather than a pair, of keypoints, which improves both precision and recall. Accordingly, we design two customized modules named cascade corner pooling and center pooling, which play the roles of enriching information collected by both top-left and bottom-right corners and providing more recognizable information at the central regions, respectively. On the MS-COCO dataset, CenterNet achieves an AP of **47.0%**, which outperforms all existing one-stage detectors by at least **4.9%**. Meanwhile, with a faster inference speed, CenterNet demonstrates quite comparable performance to the top-ranked two-stage detectors. Code is available at <https://github.com/Duankaiwen/CenterNet>.

1. Introduction

Object detection has been significantly improved and advanced with the help of deep learning, especially convolutional neural networks [11] (CNNs). In the current era, one of the most popular flowcharts is anchor-based [10, 12, 27, 31, 33], which placed a set of rectangles with pre-defined sizes, and regressed them to the desired place with the help of ground-truth objects. These approaches often need a large number of anchors to ensure a sufficiently high IoU (intersection over union) rate with the ground-truth objects, and the size and aspect ratio of each anchor box need to be manually designed. In addition, anchors are usually not

*This work was done when the first author was interning at Huawei Noah’s Ark Lab.

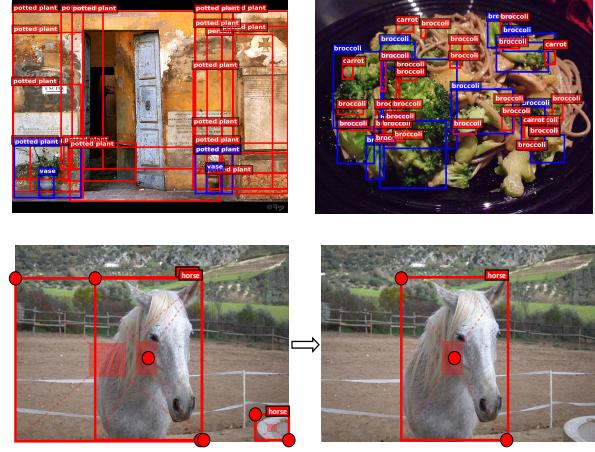


Figure 1: In the first row, we visualize the top 100 bounding boxes (according to the MS-COCO dataset standard) of CornerNet. Ground-truth and predicted objects are marked in blue and red, respectively. In the second row, we show that correct predictions can be determined by checking the central parts.

aligned with the ground-truth boxes, which is not conducive to the bounding box classification task.

To overcome the drawbacks of anchor-based approaches, a keypoint-based object detection pipeline named CornerNet [20] was proposed. It represented each object by a pair of corner keypoints, which bypassed the need of anchor boxes and achieved the state-of-the-art one-stage object detection accuracy. Nevertheless, the performance of CornerNet is still restricted by its relatively weak ability of referring to the global information of an object. That is to say, since each object is constructed by a pair of corners, the algorithm is sensitive to detect the boundary of objects, meanwhile not being aware of which pairs of keypoints should be grouped into objects. Consequently, as shown in Figure 1, it often generates some incorrect bounding boxes, most of which could be easily filtered out with complementary information, e.g., the aspect ratio.

To address this issue, we equip CornerNet with an ability of perceiving the visual patterns within each proposed region, so that it can identify the correctness of each bounding box by itself. In this paper, we present a low-cost yet effective solution named **CenterNet**, which explores the central part of a proposal, *i.e.*, the region that is close to the geometric center, with one extra keypoint. Our intuition is that, if a predicted bounding box has a high IoU with the ground-truth box, then the probability that the center keypoint in its central region is predicted as the same class is high, and vice versa. Thus, during inference, after a proposal is generated as a pair of corner keypoints, we determine if the proposal is indeed an object by checking if there is a center keypoint of the same class falling within its central region. The idea, as shown in Figure 1, is to use a triplet, instead of a pair, of keypoints to represent each object.

Accordingly, for better detecting center keypoints and corners, we propose two strategies to enrich center and corner information, respectively. The first strategy is named **center pooling**, which is used in the branch for predicting center keypoints. Center pooling helps the center keypoints obtain more recognizable visual patterns within objects, which makes it easier to perceive the central part of a proposal. We achieve this by getting out the max summed response in both horizontal and vertical directions of the center keypoint on a feature map for predicting center keypoints. The second strategy is named **cascade corner pooling**, which equips the original corner pooling module [20] with the ability of perceiving internal information. We achieve this by getting out the max summed response in both boundary and internal directions of objects on a feature map for predicting corners. Empirically, we verify that such a two-directional pooling method is more stable, *i.e.*, being more robust to feature-level noises, which contributes to the improvement of both precision and recall.

We evaluate the proposed CenterNet on the MS-COCO dataset [25], one of the most popular benchmarks for large-scale object detection. CenterNet, with both center pooling and cascade corner pooling incorporated, reports an AP of **47.0%** on the test-dev set, which outperforms all existing one-stage detectors by a large margin. With an average inference time of 270ms using a 52-layer hourglass backbone [29] and 340ms using a 104-layer hourglass backbone [29] per image, CenterNet is quite efficient yet closely matches the state-of-the-art performance of the other two-stage detectors.

The remainder of this paper is organized as follows. Section 2 briefly reviews related work, and Section 3 details the proposed CenterNet. Experimental results are given in Section 4, followed by the conclusion in Section 5.

2. Related Work

Object detection involves locating and classifying the objects. In the deep learning era, powered by deep convolutional neural networks, object detection approaches can be roughly categorized into two main types of pipelines, namely, two-stage approaches and one-stage approaches.

Two-stage approaches divide the object detection task into two stages: extract RoIs, then classify and regress the RoIs.

R-CNN [11] uses a selective search method [42] to locate RoIs in the input images and uses a DCN-based regionwise classifier to classify the RoIs independently. SPP-Net [13] and Fast-RCNN [10] improve R-CNN by extracting the RoIs from the feature maps. Faster-RCNN [33] is allowed to be trained end to end by introducing RPN (region proposal network). RPN can generate RoIs by regressing the anchor boxes. Later, the anchor boxes are widely used in the object detection task. Mask-RCNN [12] adds a mask prediction branch on the Faster-RCNN, which can detect objects and predict their masks at the same time. R-FCN [6] replaces the fully connected layers with the position-sensitive score maps for better detecting objects. Cascade R-CNN [4] addresses the problem of overfitting at training and quality mismatch at inference by training a sequence of detectors with increasing IoU thresholds. The keypoint-based object detection approaches [40, 28] are proposed to avoid the disadvantages of using anchor boxes and bounding boxes regression. Other meaningful works are proposed for different problems in object detection, *e.g.*, [47, 21] focus on the architecture design, [1, 9, 36, 44] focus on the contextual relationship, [22, 3] focus on the multi-scale unification.

One-stage approaches remove the RoI extraction process and directly classify and regress the candidate anchor boxes.

YOLO [31] uses fewer anchor boxes (divide the input image into an $S \times S$ grid) to do regression and classification. YOLOv2 [32] improves the performance by using more anchor boxes and a new bounding box regression method. SSD [27] places anchor boxes densely over an input image and use features from different convolutional layers to regress and classify the anchor boxes. DSSD [8] introduces a deconvolution module into SSD to combine low-level and high-level features. While R-SSD [17] uses pooling and deconvolution operations in different feature layers to combine low-level and high-level features. RON [19] proposes a reverse connection and an objectness prior to extract multiscale features effectively. RefineDet [45] refines the locations and sizes of the anchor boxes for two times, which inherits the merits of both one-stage and two-stage approaches. CornerNet [20] is another keypoint-based approach, which directly detects an object using a pair of corners. Although CornerNet achieves high performance, it still has more room to improve.

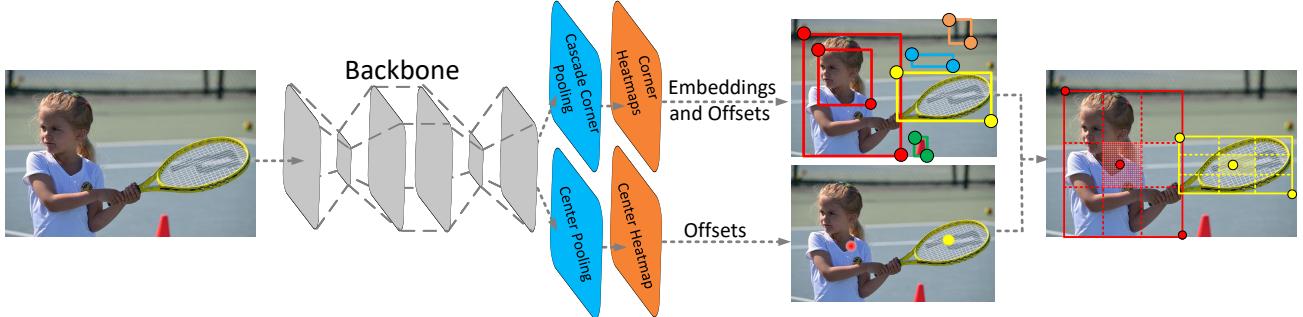


Figure 2: Architecture of CenterNet. A convolutional backbone network applies cascade corner pooling and center pooling to output two corner heatmaps and a center keypoint heatmap, respectively. Similar to CornerNet, a pair of detected corners and the similar embeddings are used to detect a potential bounding box. Then the detected center keypoints are used to determine the final bounding boxes.

Method	FD	FD ₅	FD ₂₅	FD ₅₀	FD _S	FD _M	FD _L
CornerNet	37.8	32.7	36.8	43.8	60.3	33.2	25.1

Table 1: False discovery rates (%) of CornerNet. The false discovery rate reflects the distribution of incorrect bounding boxes. The results suggest the incorrect bounding boxes account for a large proportion.

3. Our Approach

3.1. Baseline and Motivation

This paper uses CornerNet [20] as the baseline. For detecting corners, CornerNet produces two heatmaps: a heatmap of top-left corners and a heatmap of bottom-right corners. The heatmaps represent the locations of keypoints of different categories and assigns a confidence score for each keypoint. Besides, it also predicts an embedding and a group of offsets for each corner. The embeddings are used to identify if two corners are from the same object. The offsets learn to remap the corners from the heatmaps to the input image. For generating object bounding boxes, top- k left-top corners and bottom-right corners are selected from the heatmaps according to their scores, respectively. Then, the distance of the embedding vectors of a pair of corners is calculated to determine if the paired corners belong to the same object. An object bounding box is generated if the distance is less than a threshold. The bounding box is assigned a confidence score, which equals to the average scores of the corner pair.

In Table 1, we provide a deeper analysis of CornerNet. We count the FD¹ (false discovery) rate of CornerNet on the MS-COCO validation dataset, defined as the proportion of the incorrect bounding boxes. The quantitative re-

¹FD = 1 - AP, where AP denotes the average precision at IoU = [0.05 : 0.05 : 0.5] on the MS-COCO dataset. Also, FD_i = 1 - AP_i, where AP_i denotes the average precision at IoU = $i/100$, FD_{scale} = 1 - AP_{scale}, where scale = {small, medium, large}, denotes the scale of object.

sults demonstrate the incorrect bounding boxes account for a large proportion even at low IoU thresholds, e.g., CornerNet obtains 32.7% FD rate at IoU = 0.05. This means in average, 32.7 out of every 100 object bounding boxes have IoU lower than 0.05 with the ground-truth. The small incorrect bounding boxes are even more, which achieves 60.3% FD rate. One of the possible reasons lies in that CornerNet cannot look into the regions inside the bounding boxes. To make CornerNet [20] perceive the visual patterns in bounding boxes, one potential solution is to adapt CornerNet into a two-stage detector, which uses the ROI pooling [10] to look into the visual patterns in bounding boxes. However, it is known that such a paradigm is computationally expensive.

In this paper, we propose a highly efficient alternative called **CenterNet** to explore the visual patterns within each bounding box. For detecting an object, our approach uses a triplet, rather than a pair, of keypoints. By doing so, our approach is still a one-stage detector, but partially inherits the functionality of ROI pooling. Our approach only pays attention to the center information, the cost of our approach is minimal. Meanwhile, we further introduce the visual patterns within objects into the keypoint detection process by using center pooling and cascade corner pooling.

3.2. Object Detection as Keypoint Triplets

The overall network architecture is shown in Figure 2. We represent each object by a center keypoint and a pair of corners. Specifically, we embed a heatmap for the center keypoints on the basis of CornerNet and predict the offsets of the center keypoints. Then, we use the method proposed in CornerNet [20] to generate top- k bounding boxes. However, to effectively filter out the incorrect bounding boxes, we leverage the detected center keypoints and resort to the following procedure: (1) select top- k center keypoints according to their scores; (2) use the corresponding offsets to remap these center keypoints to the input image; (3) de-

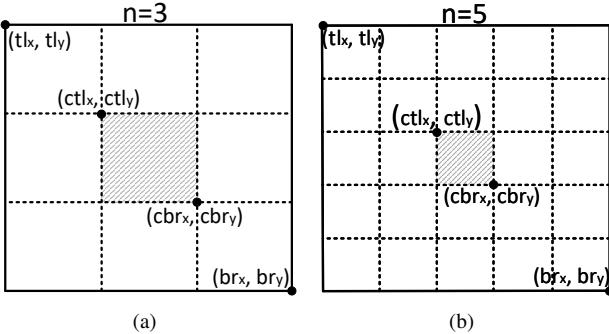


Figure 3: (a) The central region when $n = 3$. (b) The central region when $n = 5$. The solid rectangles denote the predicted bounding boxes and the shaded regions denote the central regions.

fine a central region for each bounding box and check if the central region contains center keypoints. Note that the class labels of the checked center keypoints should be same as that of the bounding box; (4) if a center keypoint is detected in the central region, we will preserve the bounding box. The score of the bounding box will be replaced by the average scores of the three points, *i.e.*, the top-left corner, the bottom-right corner and the center keypoint. If there are no center keypoints detected in its central region, the bounding box will be removed.

The size of the central region in the bounding box affects the detection results. For example, smaller central regions lead to a low recall rate for small bounding boxes, while larger central regions lead to a low precision for large bounding boxes. Therefore, we propose a scale-aware central region to adaptively fit the size of bounding boxes. The scale-aware central region tends to generate a relatively large central region for a small bounding box, while a relatively small central region for a large bounding box. Suppose we want to determine if a bounding box i needs to be preserved. Let tl_x and tl_y denote the coordinates of the top-left corner of i and br_x and br_y denote the coordinates of the bottom-right corner of i . Define a central region j . Let ctl_x and ctl_y denote the coordinates of the top-left corner of j and cbr_x and cbr_y denote the coordinates of the bottom-right corner of j . Then tl_x , tl_y , br_x , br_y , ctl_x , ctl_y , cbr_x and cbr_y should satisfy the following relationship:

$$\left\{ \begin{array}{l} ctl_x = \frac{(n+1)tl_x + (n-1)br_x}{2n} \\ ctl_y = \frac{(n+1)tl_y + (n-1)br_y}{2n} \\ cbr_x = \frac{(n-1)tl_x + (n+1)br_x}{2n} \\ cbr_y = \frac{(n-1)tl_y + (n+1)br_y}{2n} \end{array} \right. \quad (1)$$

where n is odd that determines the scale of the central re-

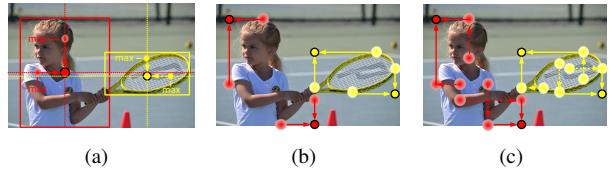


Figure 4: (a) Center pooling takes the maximum values in both horizontal and vertical directions. (b) Corner pooling only takes the maximum values in boundary directions. (c) Cascade corner pooling takes the maximum values in both boundary directions and internal directions of objects.

gion j . In this paper, n is set to be 3 and 5 for the scales of bounding boxes less and greater than 150, respectively. Figure 3 shows two central regions when $n = 3$ and $n = 5$, respectively. According to Equation (1), we can determine a scale-aware central region, then we check if the central region contains center keypoints.

3.3. Enriching Center and Corner Information

Center pooling. The geometric centers of objects do not necessarily convey very recognizable visual patterns (*e.g.*, the human head contains strong visual patterns, but the center keypoint is often in the middle of the human body). To address this issue, we propose center pooling to capture richer and more recognizable visual patterns. Figure 4(a) shows the principle of center pooling. The detailed process of center pooling is as follows: the backbone outputs a feature map, and to determine if a pixel in the feature map is a center keypoint, we need to find the maximum value in its both horizontal and vertical directions and add them together. By doing this, center pooling helps the better detection of center keypoints.

Cascade corner pooling. Corners are often outside the objects, which lacks local appearance features. CornerNet [20] uses corner pooling to address this issue. The principle of corner pooling is shown in Figure 4(b). Corner pooling aims to find the maximum values on the boundary directions so as to determine corners. However, it makes corners sensitive to the edges. To address this problem, we need to let corners “see” the visual patterns of objects. The principle of cascade corner pooling is presented in Figure 4(c). It first looks along a boundary to find a boundary maximum value, then looks inside along the location of the boundary maximum value² to find an internal maximum value, and finally, add the two maximum values together. By doing this, the corners obtain both the boundary information and the visual patterns of objects.

Both the center pooling and the cascade corner pooling can be easily achieved by combining the corner pooling [20]

²For the topmost, leftmost, bottommost and rightmost boundary, look vertically towards the bottom, horizontally towards the right, vertically towards the top and horizontally towards the left, respectively.

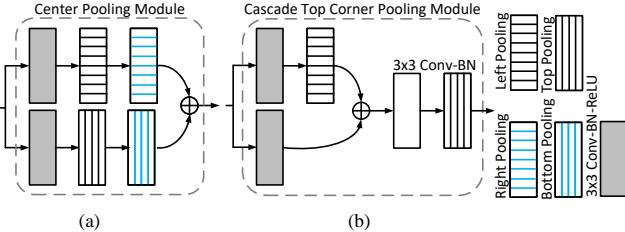


Figure 5: The structures of the center pooling module (a) and the cascade top corner pooling module (b). We achieve center pooling and the cascade corner pooling by combining the corner pooling at different directions.

at different directions. Figure 5(a) shows the structure of the center pooling module. To take a maximum value in a direction, *e.g.*, the horizontal direction, we only need to connect the left pooling and the right pooling in series. Figure 5(b) shows the structure of a cascade top corner pooling module. Compared with the top corner pooling in CornerNet [20], we add a left corner pooling before the top corner pooling.

3.4. Training and Inference

Training. Our method is implemented in Pytorch [30] and the network is trained from scratch. The resolution of the input image is 511×511 , leading to heatmaps of size 128×128 . We use the data augmentation strategy presented in [20] to train a robust model. Adam [18] is used to optimize the training loss:

$$L = L_{\text{det}}^{\text{co}} + L_{\text{det}}^{\text{ce}} + \alpha L_{\text{pull}}^{\text{co}} + \beta L_{\text{push}}^{\text{co}} + \gamma (L_{\text{off}}^{\text{co}} + L_{\text{off}}^{\text{ce}}), \quad (2)$$

where $L_{\text{det}}^{\text{co}}$ and $L_{\text{det}}^{\text{ce}}$ denote the focal losses, which are used to train the network to detect corners and center keypoints, respectively. $L_{\text{pull}}^{\text{co}}$ is a “pull” loss for corners, which is used to minimize the distance of the embedding vectors that belongs to the same objects. $L_{\text{push}}^{\text{co}}$ is a “push” loss for corners, which is used to maximize the distance of the embedding vectors that belongs to different objects. $L_{\text{off}}^{\text{co}}$ and $L_{\text{off}}^{\text{ce}}$ are ℓ_1 -losses [10], which are used to train the network to predict the offsets of corners and center keypoints, respectively. α , β and γ denote the weights for corresponding losses, which are set to 0.1, 0.1 and 1, respectively. L_{det} , L_{pull} , L_{push} and L_{off} are all defined in the CornerNet, we suggest to refer to [20] for details. We train the CenterNet on 8 Tesla V100 (32GB) GPUs and use a batch size of 48. The maximum number of iterations is 480K. We use a learning rate of 2.5×10^{-4} for the first 450K iterations and then continue training 30K iterations with a rate of 2.5×10^{-5} .

Inference. Following [20], for the single-scale testing, we input both the original and horizontally flipped images with the original resolutions into the network. While for the multi-scale testing, we input both the original and horizontally flipped images with the resolutions of 0.6, 1, 1.2, 1.5 and 1.8. We select top 70 center keypoints, top 70 top-left

corners and top 70 bottom-right corners from the heatmaps to detect the bounding boxes. We flip the bounding boxes detected in the horizontally flipped images and mix them into the original bounding boxes. Soft-nms [2] is used to remove the redundant bounding boxes. We finally select top 100 bounding boxes according to their scores as the final detection results.

4. Experiments

4.1. Dataset, Metrics and Baseline

We evaluate our method on the MS-COCO dataset [25]. It contains 80 categories and more than 1.5 million object instances. The large number of small objects makes it a very challenging dataset. We use the ‘trainval35k’ set [15] (*i.e.*, 80K training images and 35K validation images) for training and test the results on the test-dev set. We use another 5K images in the validation set to perform ablation studies and visualization experiments.

MS-COCO dataset [25] uses AP and AR metrics to characterize the performance of a detector. AP represents the average precision rate, which is computed over ten different IoU thresholds (*i.e.*, 0.5 : 0.05 : 0.95) and all categories. It is considered the single most important metric on the MS-COCO dataset. AR represents the maximum recall rate, which is computed over a fixed number of detections (*i.e.*, 1, 10 and 100) per image and averaged over all categories and the ten different IoU thresholds. Additionally, AP and AR can be used to evaluate the performance under different object scales, including small objects (area $< 32^2$), medium objects ($32^2 < \text{area} < 96^2$) and large objects (area $> 96^2$).

Our direct baseline is CornerNet [20]. Following it, we use the stacked hourglass network [29] with 52 and 104 layers as the backbone – the latter has two hourglass modules while the former has only one. All modifications on the hourglass architecture, made by [20], are preserved.

4.2. Comparisons with State-of-the-art Detectors

Table 2 shows the comparison with the state-of-the-art detectors on the MS-COCO test-dev set.

Compared with the baseline CornerNet [20], the proposed CenterNet achieves a remarkable improvement. For example, CenterNet511-52 (means that the resolution of input images is 511×511 and the backbone is Hourglass-52) reports a single-scale testing AP of 41.6%, an improvement of 3.8% over 37.8%, and a multi-scale testing AP of 43.5%, an improvement of 4.1% over 39.4%, achieved by CornerNet under the same setting. When using the deeper backbone (*i.e.*, Hourglass-104), the AP improvement over CornerNet are 4.4% (from 40.5% to 44.9%) and 4.9% (from 42.1% to 47.0%) under the single-scale and multi-scale testing, respectively. These results firmly demonstrate the effectiveness of CenterNet.

Method	Backbone	Train input	Test input	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	AR ₁	AR ₁₀	AR ₁₀₀	AR _S	AR _M	AR _L
Two-stage:															
DeNet [40]	ResNet-101 [14]	512×512	512×512	33.8	53.4	36.1	12.3	36.1	50.8	29.6	42.6	43.5	19.2	46.9	64.3
CoupleNet [47]	ResNet-101	ori.	ori.	34.4	54.8	37.2	13.4	38.1	50.8	30.0	45.0	46.4	20.7	53.1	68.5
Faster R-CNN by G-RMI [16]	Inception-ResNet-v2 [39]	~1000×600	~1000×600	34.7	55.5	36.7	13.5	38.1	52.0	-	-	-	-	-	-
Faster R-CNN +++ [14]	ResNet-101	~1000×600	~1000×600	34.9	55.7	37.4	15.6	38.7	50.9	-	-	-	-	-	-
Faster R-CNN w/ FPN [23]	ResNet-101	~1000×600	~1000×600	36.2	59.1	39.0	18.2	39.0	48.2	-	-	-	-	-	-
Faster R-CNN w/ TDM [37]	Inception-ResNet-v2	-	-	36.8	57.7	39.2	16.2	39.8	52.1	31.6	49.3	51.9	28.1	56.6	71.1
D-FCN [7]	Aligned-Inception-ResNet	~1000×600	~1000×600	37.5	58.0	-	19.4	40.1	52.5	-	-	-	-	-	-
Regionlets [43]	ResNet-101	~1000×600	~1000×600	39.3	59.8	-	21.7	43.7	50.9	-	-	-	-	-	-
Mask R-CNN [12]	ResNeXt-101	~1300×800	~1300×800	39.8	62.3	43.4	22.1	43.2	51.2	-	-	-	-	-	-
Soft-NMS [2]	Aligned-Inception-ResNet	~1300×800	~1300×800	40.9	62.8	-	23.3	43.6	53.3	-	-	-	-	-	-
Fitness R-CNN [41]	ResNet-101	512×512	1024×1024	41.8	60.9	44.9	21.5	45.0	57.5	-	-	-	-	-	-
Cascade R-CNN [4]	ResNet-101	-	-	42.8	62.1	46.3	23.7	45.5	55.2	-	-	-	-	-	-
Grid R-CNN w/ FPN [28]	ResNeXt-101	~1300×800	~1300×800	43.2	63.0	46.6	25.1	46.5	55.2	-	-	-	-	-	-
D-RFCN + SNIP (multi-scale) [38]	DPN-98 [5]	~2000×1200	~2000×1200	45.7	67.3	51.1	29.3	48.8	57.1	-	-	-	-	-	-
PANet (multi-scale) [26]	ResNeXt-101	~1400×840	~1400×840	47.4	67.2	51.8	30.1	51.7	60.0	-	-	-	-	-	-
One-stage:															
YOLOv2 [32]	DarkNet-19	544×544	544×544	21.6	44.0	19.2	5.0	22.4	35.5	20.7	31.6	33.3	9.8	36.5	54.4
DSOD300 [34]	DS/64-192-48-1	300×300	300×300	29.3	47.3	30.6	9.4	31.5	47.0	27.3	40.7	43.0	16.7	47.1	65.0
GRP-DSOD320 [35]	DS/64-192-48-1	320×320	320×320	30.0	47.9	31.8	10.9	33.6	46.3	28.0	42.1	44.5	18.8	49.1	65.0
SSD513 [27]	ResNet-101	513×513	513×513	31.2	50.4	33.3	10.2	34.5	49.8	28.3	42.1	44.4	17.6	49.2	65.8
DSSD513 [8]	ResNet-101	513×513	513×513	33.2	53.3	35.2	13.0	35.4	51.1	28.9	43.5	46.2	21.8	49.1	66.4
RefineDet512 (single-scale) [45]	ResNet-101	512×512	512×512	36.4	57.5	39.5	16.6	39.9	51.4	-	-	-	-	-	-
CornerNet511 (single-scale) [20]	Hourglass-52	511×511	ori.	37.8	53.7	40.1	17.0	39.0	50.5	33.9	52.3	57.0	35.0	59.3	74.7
RetinaNet800 [24]	ResNet-101	800×800	800×800	39.1	59.1	42.3	21.8	42.7	50.2	-	-	-	-	-	-
CornerNet511 (multi-scale) [20]	Hourglass-52	511×511	≤1.5×	39.4	54.9	42.3	18.9	41.2	52.7	35.0	53.5	57.7	36.1	60.1	75.1
CornerNet511 (single-scale) [20]	Hourglass-104	511×511	ori.	40.5	56.5	43.1	19.4	42.7	53.9	35.3	54.3	59.1	37.4	61.9	76.9
RefineDet512 (multi-scale) [45]	ResNet-101	512×512	≤2.25×	41.8	62.9	45.7	25.6	45.1	54.1	-	-	-	-	-	-
CornerNet511 (multi-scale) [20]	Hourglass-104	511×511	≤1.5×	42.1	57.8	45.3	20.8	44.8	56.7	36.4	55.7	60.0	38.5	62.7	77.4
CenterNet511 (single-scale)	Hourglass-52	511×511	ori.	41.6	59.4	44.2	22.5	43.1	54.1	34.8	55.7	60.1	38.6	63.3	76.9
CenterNet511 (single-scale)	Hourglass-104	511×511	ori.	44.9	62.4	48.1	25.6	47.4	57.4	36.1	58.4	63.3	41.3	67.1	80.2
CenterNet511 (multi-scale)	Hourglass-52	511×511	≤1.8×	43.5	61.3	46.7	25.3	45.3	55.0	36.0	57.2	61.3	41.4	64.0	76.3
CenterNet511 (multi-scale)	Hourglass-104	511×511	≤1.8×	47.0	64.5	50.7	28.9	49.9	58.9	37.5	60.3	64.8	45.1	68.3	79.7

Table 2: Performance comparison (%) with the state-of-the-art methods on the MS-COCO test-dev dataset. CenterNet outperforms all existing one-stage detectors by a large margin and ranks among the top of state-of-the-art two-stage detectors.

Meanwhile, it can be seen that the most contribution comes from the small objects. For instance, CenterNet511-52 improves the AP for small objects by 5.5% (single-scale) and by 6.4% (multi-scale). As for the backbone Hourglass-104, the improvements are 6.2% (single-scale) and by 8.1% (multi-scale), respectively. The benefit stems from the center information modeled by the center keypoints: the smaller the scale of an incorrect bounding box is, the lower probability a center keypoint can be detected in its central region. Figure 6(a) and Figure 6(b) show some qualitative comparisons, which demonstrate the effectiveness of CenterNet in reducing small incorrect bounding boxes.

CenterNet also leads to a large improvement for reducing medium and large incorrect bounding boxes. As Table 2 shows, CenterNet511-104 improves the single-scale testing AP by 4.7% (from 42.7% to 47.4%) and 3.5% (from 53.9% to 57.4%), respectively. Figure 6(c) and Figure 6(d) show some qualitative comparisons for reducing medium and large incorrect bounding boxes. It is worth noting that the AR is also significantly improved, with the best performance achieved with multi-scale testing. This is because our approach removes lots of incorrect bounding boxes, which is equivalent to improving the confidence of those bounding boxes with accurate locations but lower scores.

When comparing other one-stage approaches, CenterNet511-52 reports 41.6% single-scale testing AP. This achievement is already better than those using deeper models (*e.g.*, RetinaNet800 [24] and RefineDet [45]). The best performance of CenterNet is AP **47.0%**, dramatically

surpassing all the published one-stage approaches to our best knowledge.

At last, one can observe that the performance of CenterNet is also competitive with the two-stage approaches, *e.g.*, the single-scale testing AP of CenterNet511-52 is comparable to the two-stage approach Fitness R-CNN [41] (41.6% vs. 41.8%) and that of CenterNet511-104 is comparable to D-RFCN + SNIP [38] (44.9% vs. 45.7%), respectively. Nevertheless, it should be mentioned that two-stage approaches usually use larger resolution input images (*e.g.*, ~1000 × 600), which significantly improves the detection accuracy especially for small objects. The multi-scale testing AP **47.0%** achieved by CenterNet511-104 closely matches the state-of-the-art AP 47.4%, achieved by the two-stage detector PANet [26]. We present some qualitative detection results in Figure 7.

4.3. Incorrect Bounding Box Reduction

The AP [25] metric reflects how many high quality object bounding boxes (usually $\text{IoU} \geq 0.5$) a network can predict, but cannot directly reflect how many incorrect object bounding boxes (usually $\text{IoU} \ll 0.5$) a network generates. The FD rate is a suitable metric, which reflects the proportion of the incorrect bounding boxes. Table 3 shows the FD rates for CornerNet and CenterNet. CornerNet generates many incorrect bounding boxes even at $\text{IoU} = 0.05$ threshold, *i.e.*, CornerNet511-52 and CornerNet511-104 obtain 35.2% and 32.7% FD rate, respectively. On the other hand, CenterNet generates more small incorrect bounding

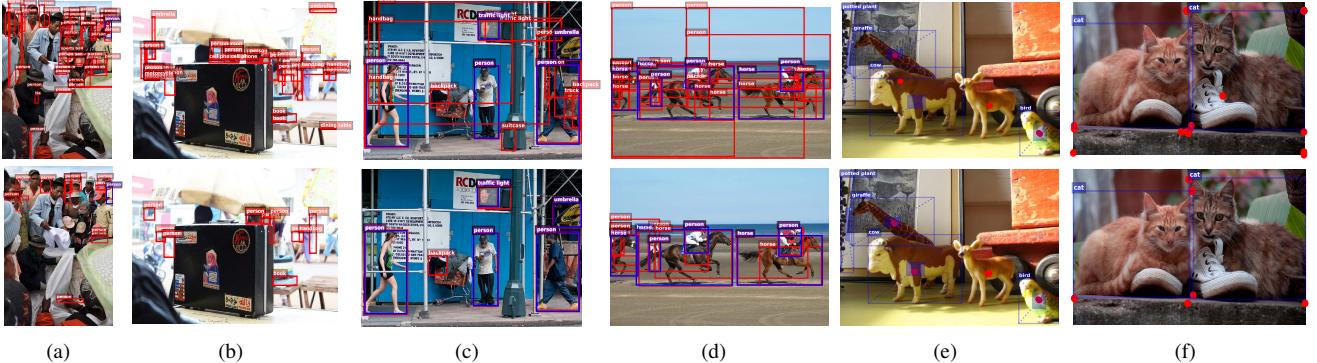


Figure 6: (a) and (b) show the small incorrect bounding boxes are significantly reduced by modeling center information. (c) and (d) show that the center information works for reducing medium and large incorrect bounding boxes. (e) shows the results of detecting the center keypoints without/with the center pooling. (f) shows the results of detecting the corners with corner pooling and cascade corner pooling, respectively. The blue boxes above denote the ground-truth. The red boxes and dots denote the predicted bounding boxes and keypoints, respectively.



Figure 7: Some qualitative detection results on the MS-COCO validation dataset. Only detections with scores higher than 0.5 are shown.

Method	FD	FD ₅	FD ₂₅	FD ₅₀	FD _S	FD _M	FD _L
CornerNet511-52	40.4	35.2	39.4	46.7	62.5	36.9	28.0
CenterNet511-52	35.1	30.7	34.2	40.8	53.0	31.3	24.4
CornerNet511-104	37.8	32.7	36.8	43.8	60.3	33.2	25.1
CenterNet511-104	32.4	28.2	31.6	37.5	50.7	27.1	23.0

Table 3: Comparison of false discovery rates (%) of CornerNet and CenterNet on the MS-COCO validation dataset. The results suggest CenterNet avoids a large number of incorrect bounding boxes, especially for small incorrect bounding boxes.

boxes than medium and large incorrect bounding boxes, which reports 62.5% for CornerNet511-52 and 60.3% for CornerNet511-104, respectively. Our CenterNet decreases the FD rates at all criteria via exploring central regions. For instance, CenterNet511-52 and CenterNet511-104 decrease FD₅ by both 4.5%. In addition, the FD rates for small bounding boxes decrease the most, which are 9.5% by CenterNet511-52 and 9.6% by CenterNet511-104, respectively. This is also the reason why the AP improvement for small objects is more prominent.

4.4. Inference Speed

The proposed CenterNet explores the visual patterns within each proposed region with minimal costs. To ensure a fair comparison, we test the inference speed of both CornerNet [20] and CenterNet on a NVIDIA Tesla P100 GPU. We obtain that the average inference time of CornerNet511-

104 is 300ms per image and that of CenterNet511-104 is 340ms. Meanwhile, using the Hourglass-52 backbone can speed up the inference speed. Our CenterNet511-52 takes an average of 270ms to process per image, which is faster and more accurate than CornerNet511-104.

4.5. Ablation Study

Our work has contributed three components, including central region exploration, center pooling and cascade corner pooling. To analyze the contribution of each individual component, an ablation study is given here. The baseline is CornerNet511-52 [20]. We add the three components to the baseline one by one and follow the default parameter setting detailed in Section 4.1. The results are given in Table 4.

Central region exploration. To understand the importance of the central region exploration (see CRE in the table), we add a center heatmap branch to the baseline and use a triplet of keypoints to detect bounding boxes. For the center keypoint detection, we only use conventional convolutions. As presented in the third row in Table 4, we improve the AP by 2.3% (from 37.6% to 39.9%). However, we find that the improvement for the small objects (that is 4.6%) is more significant than that for other object scales. The improvement for large objects is almost negligible (from 52.2% to 52.3%). This is not surprising because, from a probabilistic point of view, the center keypoint for a small object is easier

CRE	CTP	CCP	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L	AR ₁	AR ₁₀	AR ₁₀₀	AR _S	AR _M	AR _L
			37.6	53.3	40.0	18.5	39.6	52.2	33.7	52.2	56.7	37.2	60.0	74.0
	✓		38.3	54.2	40.5	18.6	40.5	52.2	34.0	53.0	57.9	36.6	60.8	75.8
✓			39.9	57.7	42.3	23.1	42.3	52.3	33.8	54.2	58.5	38.7	62.4	74.4
✓	✓		40.8	58.6	43.6	23.6	43.6	53.6	33.9	54.5	59.0	39.0	63.2	74.7
✓	✓	✓	41.3	59.2	43.9	23.6	43.8	55.8	34.5	55.0	59.2	39.1	63.5	75.1

Table 4: Ablation study on the major components of CenterNet511-52 on the MS-COCO validation dataset. The CRE denotes central region exploration, the CTP denotes center pooling, and the CCP denotes cascade corner pooling.

Method	AP	AP ₅₀	AP ₇₅	AP _S	AP _M	AP _L
CenterNet511-52 w/o GT	41.3	59.2	43.9	23.6	43.8	55.8
CenterNet511-52 w/ GT	56.5	78.3	61.4	39.1	60.3	70.3
CenterNet511-104 w/o GT	44.8	62.4	48.2	25.9	48.9	58.8
CenterNet511-104 w/ GT	58.1	78.4	63.9	40.4	63.0	72.1

Table 5: Error analysis of center keypoints via using ground-truth. we replace the predicted center keypoints with the ground-truth values, the results suggest there is still room for improvement in detecting center keypoints.

to be located than that of a large object.

Center pooling. To demonstrate the effectiveness of proposed center pooling, we then add the center pooling module to the network (see CTP in the table). The fourth row in Table 4 shows that center pooling improves the AP by 0.9% (from 39.9% to 40.8%). Notably, with the help of center pooling, we improve the AP for large objects by 1.4% (from 52.2% to 53.6%), which is much higher than the improvement using conventional convolutions (*i.e.*, 1.4% vs. 0.1%). It demonstrates that our center pooling is effective in detecting center keypoints of objects, especially for large objects. Our explanation is that center pooling can extract richer internal visual patterns, and larger objects contain more accessible internal visual patterns. Figure 6(e) shows the results of detecting center keypoints without/with center pooling. We can see the conventional convolution fails to locate the center keypoint for the cow, but with center pooling, it successfully locates the center keypoint.

Cascade corner pooling. We replace corner pooling [20] with cascade corner pooling to detect corners (see CCP in the table). The second row in Table 4 shows the results that we test on the basis of CornerNet511-52. We find that cascade corner pooling improves the AP by 0.7% (from 37.6% to 38.3%). The last row shows the results that we test on the basis of CenterNet511-52, which improves the AP by 0.5% (from 40.8% to 41.3%). The results of the second row show there is almost no change in the AP for large objects (*i.e.*, 52.2% vs. 52.2%), but the AR is improved by 1.8% (from 74.0% to 75.8%). This suggests that cascade corner pooling can “see” more objects due to the rich internal visual patterns in large objects, but too rich visual patterns may interfere with its perception for the boundary information, leading to many inaccurate bounding boxes. After equipping with our CenterNet, the inaccurate bound-

ing boxes are effectively suppressed, which improves the AP for large objects by 2.2% (from 53.6% to 55.8%). Figure 6(f) shows the result of detecting corners with corner pooling or cascade corner pooling. We can see that cascade corner pooling can successfully locate a pair of corners for the cat on the left while corner pooling cannot.

4.6. Error Analysis

The exploration of visual patterns within each bounding box depends on the center keypoints. In other words, once a center keypoint is missed, the proposed CenterNet would miss the visual patterns within the bounding box. To understand the importance of center keypoints, we replace the predicted center keypoints with the ground-truth values and evaluate performance on the MS-COCO validation dataset. Table 5 shows that using the ground-truth center keypoints improves the AP from 41.3% to 56.5% for CenterNet511-52 and from 44.8% to 58.1% for CenterNet511-104, respectively. APs for small, medium and large objects are improved by 15.5%, 16.5%, and 14.5% for CenterNet511-52 and 14.5%, 14.1%, and 13.3% for CenterNet511-104, respectively. This demonstrates that the detection of center keypoints is far from the bottleneck.

5. Conclusions

In this paper, we propose CenterNet, which detects objects using a triplet, including one center keypoint and two corners. Our approach addresses the problem that CornerNet lacks an additional look into the cropped regions by exploring the visual patterns within each proposed region with minimal costs. In fact, this is a common defect for all one-stage approaches. As one-stage approaches remove the RoI extraction process, they cannot pay attention to internal information within cropped regions.

An intuitive explanation of our contribution lies in that we equip a one-stage detector with the ability of two-stage approaches, with an efficient discriminator being added. We believe that our idea of adding an extra branch for the center keypoint can be potentially generalized to other existing one-stage approaches (*e.g.*, SSD [27]). Meanwhile, some advanced training strategies [46] can be used for better performance. We leave as our future work.

References

- [1] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2874–2883, 2016.
- [2] N. Bodla, B. Singh, R. Chellappa, and L. 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] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision*, pages 354–370. Springer, 2016.
- [4] Z. Cai and N. Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.
- [5] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng. Dual path networks. In *Advances in neural information processing systems*, pages 4467–4475, 2017.
- [6] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.
- [7] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.
- [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.
- [9] S. Gidaris and N. Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *Proceedings of the IEEE international conference on computer vision*, pages 1134–1142, 2015.
- [10] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
- [11] R. Girshick, J. Donahue, T. Darrell, and J. 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.
- [12] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [13] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.
- [14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [15] D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *European conference on computer vision*, pages 340–353. Springer, 2012.
- [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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7310–7311, 2017.
- [17] J. Jeong, H. Park, and N. Kwak. Enhancement of ssd by concatenating feature maps for object detection. *arXiv preprint arXiv:1705.09587*, 2017.
- [18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *Computer science*, 2014.
- [19] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen. Ron: Reverse connection with objectness prior networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5936–5944, 2017.
- [20] H. Law and J. Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision*, pages 734–750, 2018.
- [21] H. Lee, S. Eum, and H. Kwon. Me r-cnn: Multi-expert r-cnn for object detection. *arXiv preprint arXiv:1704.01069*, 2017.
- [22] Y. Li, Y. Chen, N. Wang, and Z. Zhang. Scale-aware trident networks for object detection. *arXiv preprint arXiv:1901.01892*, 2019.
- [23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
- [24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [25] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.
- [26] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.
- [27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- [28] X. Lu, B. Li, Y. Yue, Q. Li, and J. Yan. Grid r-cnn. 2018.
- [29] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.
- [30] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.
- [31] J. Redmon, S. Divvala, R. Girshick, and A. 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.

- [32] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.
- [33] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [34] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue. Dsod: Learning deeply supervised object detectors from scratch. In *Proceedings of the IEEE international conference on computer vision*, pages 1919–1927, 2017.
- [35] Z. Shen, H. Shi, R. Feris, L. Cao, S. Yan, D. Liu, X. Wang, X. Xue, and T. S. Huang. Learning object detectors from scratch with gated recurrent feature pyramids. *arXiv preprint arXiv:1712.00886*, 2017.
- [36] A. Shrivastava and A. Gupta. Contextual priming and feedback for faster r-cnn. In *European conference on computer vision*, pages 330–348, 2016.
- [37] A. Shrivastava, R. Sukthankar, J. Malik, and A. Gupta. Beyond skip connections: Top-down modulation for object detection. *arXiv preprint arXiv:1612.06851*, 2016.
- [38] B. Singh and L. S. Davis. An analysis of scale invariance in object detection snip. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3578–3587, 2018.
- [39] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI conference on artificial intelligence*, 2017.
- [40] L. Tychsen-Smith and L. 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] L. Tychsen-Smith and L. 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] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.
- [43] H. Xu, X. Lv, X. Wang, Z. Ren, N. Bodla, and R. Chellappa. Deep regionlets for object detection. In *Proceedings of the European conference on computer vision*, pages 798–814, 2018.
- [44] X. Zeng, W. Ouyang, B. Yang, J. Yan, and X. Wang. Gated bi-directional cnn for object detection. In *European conference on computer vision*, pages 354–369. Springer, 2016.
- [45] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. Single-shot refinement neural network for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4203–4212, 2018.
- [46] R. Zhu, S. Zhang, X. Wang, L. Wen, H. Shi, L. Bo, and T. Mei. Scratchdet: Training single-shot object detectors from scratch. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019.
- [47] Y. Zhu, C. Zhao, J. Wang, X. Zhao, Y. Wu, and H. Lu. Coupletent: Coupling global structure with local parts for object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 4126–4134, 2017.