CenterNet++ for Object Detection

Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Member, IEEE, Qingming Huang, Fellow, IEEE, and Qi Tian, Fellow, IEEE

Abstract—There are two mainstream approaches for object detection: top-down and bottom-up. The state-of-the-art approaches are mainly top-down methods. In this paper, we demonstrate that bottom-up approaches show competitive performance compared with top-down approaches and have higher recall rates. Our approach, named CenterNet, detects each object as a triplet of keypoints (top-left and bottom-right corners and the center keypoint). We first group the corners according to some designed cues and confirm the object locations based on the center keypoints. The corner keypoints allow the approach to detect objects of various scales and shapes and the center keypoint reduces the confusion introduced by a large number of false-positive proposals. Our approach is an anchor-free detector because it does not need to define explicit anchor boxes. We adapt our approach to backbones with different structures, including ‘hourglass’-like networks and ‘pyramid’-like networks, which detect objects in single-resolution and multi-resolution feature maps, respectively. On the MS-COCO dataset, CenterNet with Res2Net-101 and Swin-Transformer achieve average precisions (APs) of 53.7% and 57.1%, respectively, outperforming all existing bottom-up detectors and achieving state-of-the-art performance. We also design a real-time CenterNet model, which achieves a good trade-off between accuracy and speed, with an AP of 43.6% at 30.5 frames per second (FPS).

Index Terms—Anchor-free, bottom-up, deep learning, object detection.

I. INTRODUCTION

There are currently two main types of object detection methods: bottom-up approaches [12], [30], [31], [69] and top-down approaches [9], [10], [38], [46]. Many researchers believe that bottom-up approaches are time-consuming and introduce more false positives, and top-down approaches have gradually become the mainstream approaches due to their effectiveness in practice. Top-down approaches model each object as a prior or a predefined anchor box, and predict the corresponding offsets to the bounding box. Top-down approaches can perceive whole objects, which simplifies the postprocessing steps in generating bounding boxes. However, they usually suffer from difficulty in perceiving objects with peculiar shapes (e.g., objects with large aspect ratios). For example, CornerNet [30], a representative bottom-up approach, models each object using a pair of corner keypoints and achieves state-of-the-art detection accuracy. Nevertheless, the performance of CornerNet is still restricted by its relatively weak ability to determine the global information of an object; that is, because each object is constructed according to a pair of corner keypoints, the algorithm sensitively detects the boundaries of objects without being aware of which pairs of keypoints should be grouped into objects. Consequently, as shown in Fig. 1(b), CornerNet often generates incorrect bounding boxes, most of which could be

![Fig. 1. (a) Top-down approach Faster R-CNN [48] often fails to precisely locate objects with a peculiar shapes (such as those with extreme aspect ratios). The blue and red bounding boxes indicate false positives and false negatives, respectively. (b) We visualize the top 100 bounding boxes (according to the MS-COCO dataset standard) of CornerNet [30], which shows a large number of false positives. The blue and red bounding boxes indicate true positives and false positives, respectively. (c) The concept of the proposed CenterNet. We show that correct predictions can be determined by verifying the central keypoints.](image-url)
easily filtered out with some complementary information, e.g., the aspect ratio.

Driven by the analysis of bottom-up approaches, our hypothesis is that **bottom-up approaches are competitive with top-down approaches if their ability to perceive the global information of objects is improved**. In this paper, we present a low-cost yet effective solution named CenterNet, a strong bottom-up object detection approach that detects each object according to triplet keypoints (top-left and bottom-right corners and the center). CenterNet explores the central part of a proposal, i.e., the region that is close to the geometric center of a box, with an additional keypoint compared to conventional approaches. We hypothesize that if a predicted bounding box has a high intersection over union (IoU) with the ground-truth box, the probability that the center keypoint in the central region of the bounding box will be 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 verifying whether a center keypoint of the same class falls within the central region of the proposal. This concept is shown in Fig. 1(c).

We design two frameworks to adapt to networks with different structures. The first framework is developed for ‘hourglass’-like networks, which detect objects in single-resolution feature maps. Hourglass [43] networks are very popular for performing keypoint estimation tasks, and we apply this type of network to better predict the corners and center keypoints. We also design our framework to fit ‘pyramid’-like networks, which detect objects in multi-resolution feature maps. This approach has two main advantages: stronger generality, as most of networks have ‘pyramid’-like structures, e.g., ResNet [25] and its variants; and higher detection accuracy, as objects with different scales are detected in various receptive fields. Although the pyramid structure has been widely applied in the top-down approaches, to the best of our knowledge, this is the first time that is has been used in bottom-up approaches.

We evaluate the proposed CenterNet on the MS-COCO dataset [37], one of the most popular benchmark datasets for large scale object detection. CenterNet, with Res2Net-101 [18] and Swin-Transformer [39], achieves APs of 53.7 and 57.1, respectively, outperforming all existing bottom-up detectors by a large margin. We also design a real-time CenterNet, which achieves a good trade-off between accuracy and speed with an AP of 43.6 at 30.5 FPS. CenterNet is efficient yet closely matches the state-of-the-art performance of existing top-down approaches.

A preliminary version of this paper was published in [13]. In this extended version, we improve the work based on the following aspects. (i) The original CenterNet was applied only with an Hourglass network [43] as the backbone, in which all the objects are detected only in single-resolution feature maps. We extend the idea of CenterNet to networks with pyramid structures, thereby allowing CenterNet to detect objects in multi-resolution feature maps. To this end, we propose new methods for detecting keypoints (including corners and center keypoints) and grouping keypoints. (ii) In this version, due to the new design of CenterNet, we investigate more backbones with the pyramid structures, including ResNet [25], ResNext [61] and Res2Net [18]. Moreover, we even report the detection results using the Transformer [39] as a backbone. The experimental results show that the detection accuracy is significantly improved by introducing the pyramid structure, which allows the network to use richer receptive fields to detect objects. (iii) We present a real-time CenterNet that achieves a better accuracy/speed trade-off than popular detectors.

The main contributions of this work can be summarized as follows:

- We propose a strong bottom-up object detection approach named CenterNet. CenterNet detects each object as a triplet keypoint and can thus locate objects objects with arbitrary geometries and perceive the global information of objects.
- We design two frameworks to adapt to networks with different structures, which improves the generalizability of our method. Thus, our approach is applicable to essentially all networks.
- CenterNet achieves state-of-the-art detection accuracy among bottom-up approaches and closely matches the state-of-the-art performance of existing top-down approaches.
- With properly reduced structural complexity, CenterNet achieves a satisfying trade-off between accuracy and speed. Thus, we demonstrate that bottom-up approaches are necessary and competitive with top-down approaches.

The remainder of this paper is organized as follows. Section II briefly reviews related work, and Section III presents the details of the proposed CenterNet. The experimental results are provided in Section IV, and the conclusions are discussed in Section V.

## II. RELATED WORK

Object detection involves locating and classifying objects. In deep learning, with the development of deep convolutional neural networks, object detection approaches have made a great progress. Here, we follow the convention [69], [72] in using ‘top-down’ and ‘bottom-up’ to indicate how the detection algorithm locate objects – the ‘top-down’ approaches locate objects by placing a number of anchors beyond pre-defined keypoints or grids, and the ‘bottom-up’ approaches instead organize the keypoints into objects.

**Top-down approaches** first find the proposals that represent the whole objects, and then determine the classes and the bounding boxes of the objects by classifying and regressing the proposals. The proposals could be further divided into anchor-based and anchor-free according to the different forms of proposals.

**Anchor-based** proposals, which are also called anchors, start with rectangles that have different predefined sizes, scales and shapes. They are uniformly distributed in the feature maps and are trained to regress to the desired location with the help of ground-truth objects. Some approaches focus more on the quality of the detection results. One of the most representative approaches is R-CNN [22]. This approach divides the process of object determination into two stages. Some meaningful proposals are selected in the first stage, and verified in...
the second stage. Many works have been proposed based on expansions of R-CNN, such as SPPNet [24], Fast R-CNN [21], Faster R-CNN [48], Cascade R-CNN [3], MR-CNN [20], ION [1], OHEM [52], HyperNet [29], CRAFT [62], R-FCN [9], FPN [35], Libra R-CNN [44], Mask R-CNN [23], FitnessNMS [58], Grid R-CNN [41], TridentNet [34], etc. In contrast, other approaches focus on the detection speed. These approaches usually do not have a proposal verification stage. The representative approaches include SSD [38], SSD [17], RON [28], YOLOv2 [46], RetinaNet [36], RefineDet [65], AlignDet [7], ATSS [64], M2Det [67], GF [33], FreeAnchor [66], FSAF [71], etc.

Despite the great success of the application of anchors, they have several drawbacks, e.g., a large number of anchors are often required to ensure a sufficiently high intersection over union (IoU) rate with the ground-truth objects, and the size and aspect ratio of each anchor box need to be manually designed. Therefore, very neat anchor-free proposals have been proposed. These anchor-free proposals do not use anchors and instead represent objects according to points within the objects. The key to the anchor-free proposals is accurately predicting the labels of the relative sparse points and the distances from the points to the object borders. Typical approaches include the YOLO series [19], [45], FCOS [56], Objects as Points [68], FoveaBox [27], SAPD [70], and the RepPoints series [8], [63], etc.

**Bottom-up approaches** detect the individual parts of objects instead of perceiving the objects as a whole. Subsequently, the individual parts that belong to the same object are grouped together by some trainable postprocessing algorithms. The bottom-up approach was first proposed before the deep learning era. Felzenszwalb et al. represented objects using mixtures of multiscale deformable part models, which are known as DPMs [16]. Recently, keypoint estimation [42] has inspired a new type of object detection approach, in which objects are recognized by detecting and grouping keypoints. For instance, CornerNet [30] and CornerNet-lite [31] detect objects as paired corners, while ExtremeNet [69] detects four extreme points (top-left, top-right, bottom-left, bottom-right) of an object. Since bottom-up approaches do not need anchors, they are anchor-free detectors. Most bottom-up approaches are based on state-of-the-art keypoint estimation frameworks [4], [60], which have some drawbacks, e.g., they rely too much on high-resolution heatmap and have slow inference speed.

### III. OUR APPROACH

#### A. A Drawback of Top-Down Approaches

We note that there are two important classes of object detection approaches: top-down and bottom-up. Based on the discussions in the above sections, we suggest that **bottom-up approaches have better potential in locating objects with arbitrary geometries and thus may achieve higher recall scores**. Most top-down approaches are anchor-based, which are very empirical (e.g., to improve efficiency, only anchors with common sizes and aspect ratios are considered); moreover, their shapes and locations are relatively fixed, although the subsequent bounding-box regression process could slightly change their states. Therefore, the detectors tend to miss objects with peculiar shapes. Fig. 1(a) shows a typical example of a missed detection. We also provide a quantitative study, as shown in Table I. Three representative approaches and our work are evaluated on the MS-COCO validation dataset. Table I shows that the top-down approaches obtain significantly lower recall scores than the bottom-up approaches, especially for objects with peculiar geometries, e.g., objects with scales larger than 300\(^2\) pixels or aspect ratios larger than 5 : 1. This result is not surprising because, on the one hand, Faster R-CNN [48], a typical anchor-based top-down approach, there are no predefined anchors that can match these objects. On the other hand, FCOS [56], a typical anchor-free top-down approach, has difficulty accurately regressing the long distances between the border and the proposal. Since bottom-up approaches usually detect the individual parts of objects and group them into objects, this problem is somewhat reduced. Moreover, we report the results of the proposed CenterNet, demonstrating that CenterNet has the advantages of bottom-up approaches for flexibly locating objects, especially those with peculiar geometries. Although bottom-up approaches have high recall rates, they often generate many false positives. Taking CornerNet [30] as an example, two heatmaps are produced for detecting corners: a heatmap of top-left corners and a heatmap of bottom-right corners. The heatmaps represent the locations of keypoints for different categories and are used to assign a confidence score for each keypoint. In addition, CornerNet predicts an embedding and a group of offsets for each corner (as shown in Fig. 2). These embeddings are used to identify if two corners are from the same object. The offsets are used to remap the corners from the heatmaps to the input image. To generate object bounding boxes, the top-\(k\) left-top corners and bottom-right corners are selected from the heatmaps according to their scores. Then, the distance between 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 specified threshold. The bounding box is assigned a confidence score, which is equal to the average scores of the corner pair.

In Table II, we provide a deeper analysis of the performance of CornerNet. We calculate the average false discovery (AF\(^2\)) rate of CornerNet based on the MS-COCO validation dataset, which is defined as the proportion of incorrect bounding boxes. The quantitative results demonstrate that incorrect bounding boxes account for a large proportion of the total number of bounding boxes, even at low IoU thresholds, e.g., CornerNet obtains an AF scores of 32.7% at an IoU of 0.05. This result suggests that on average, 32.7 of every 100 object bounding boxes have an IoU with the ground-truth lower than 0.05. There are even more small incorrect bounding boxes, with an AF score of 60.3% AF.

\(^1\)AF = 1 − AP, where AP denotes the average precision at IoU = [0.05 : 0.05 : 0.5] based on the MS-COCO dataset. Additionally, \(AF = 1 − AP\), \(AP\) denotes the average precision at IoU = \(i/100\), and \(AF_{\text{scale}} = 1 − AP_{\text{scale}}\), where \(scale = \{\text{small}, \text{medium}, \text{large}\}\), denotes the scale of the object.
TABLE I

| Method          | Backbone     | AR  | AR₁⁺ | AR₂⁺ | AR₃⁺ | AR₄⁺ | AR₅⁻ | AR₆⁺ | AR₇⁻ | AR₈⁺ |
|-----------------|--------------|-----|------|------|------|------|------|------|------|------|
| Top-down:       | X-101-64x4d | 57.6| 73.8 | 77.5 | 79.2 | 86.2 | 43.8 | 43.0 | 34.3 | 23.2 |
| Faster R-CNN [48] | X-101-64x4d | 64.9| 82.3 | 87.9 | 89.8 | 95.0 | 45.5 | 40.8 | 34.1 | 23.4 |
| FCOS [56]       | X-101-64x4d |     |      |      |      |      |      |      |      |      |
| Bottom-up       |              |     |      |      |      |      |      |      |      |      |
| CornerNet [30]  | HG-104       | 66.8| 85.8 | 92.6 | 95.5 | 98.5 | 50.1 | 48.3 | 40.4 | 36.5 |
| CenterNet (this work) | HG-104 | 66.8| 87.1 | 93.2 | 95.2 | 96.9 | 50.7 | 45.6 | 40.1 | 32.3 |

In this experiment, we report the AR computed by targets with different aspect ratios and sizes. To eliminate the influence of other factors on the AR, we exclude the impacts of bounding-box categories and scores on the recall and compute the AR by allowing at most 1000 object proposals. AR₁⁺, AR₂⁺, AR₃⁺ and AR₄⁺ denote box area in the ranges of [96², 20²], [20², 30²], [30², 40²], and [40², ∞), respectively. ‘X’ and ‘HG’ denote ResNetX and Hourglass, respectively.

TABLE II

| Method          | AF  | AF₂ | AF₃ | AF₅ | AF₇ | AF₁₀ | AF₂₀ | AF₃₀ | AF₅₀ | AF₁₀₀ |
|-----------------|-----|-----|-----|-----|-----|------|------|------|------|-------|
| CornerNet       | 37.8| 32.7| 36.5| 43.8| 60.3| 33.2 | 25.1 |

The AF score reflects the number of incorrect bounding boxes (false positives). The results suggest that false positives account for a large proportion of the results.

CornerNet [30] to perceive the visual patterns in the bounding boxes, one potential solution is to adapt CornerNet into a two-stage detector that uses RoI pooling [21] to perceive the visual patterns in the bounding boxes. However, these paradigms are computationally expensive.

In this paper, we propose a highly efficient alternative called CenterNet to explore the visual patterns within each bounding box. To detect an object, our approach uses a triplet, rather than a pair, of keypoints. As a result, our approach focuses on the center information, while maintaining the minimal computational costs, and inheriting the functionality of RoI pooling. Furthermore, we design two frameworks for detecting objects in single-resolution feature maps and multisolution feature maps. The former is applied to the keypoint estimation network to improve the performance in detecting the corners and center keypoints. The latter is more popular in the object detection task because it has better generality and obtains richer detection receptive fields. The two frameworks have slightly different designs, and we provide detailed explanations in the next subsection.

B. Object Detection as Keypoint Triplets

1) Single-Resolution Detection Framework: Inspired by pose estimation, we apply the networks that are commonly used in pose estimation to better detect the corners and center keypoints, most of which detect the keypoints in single-resolution feature maps, e.g., hourglass networks [43]. The overall network architecture is shown in Fig. 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 [30] to perceive the visual patterns in the bounding boxes and predict the offsets of the center keypoints. Then, we use the method applied in CornerNet [30] to generate the top-k bounding boxes. To effectively identify the incorrect bounding boxes, we use the detected center keypoints and the following procedure: (1) Select the top-k center keypoints according to their scores. (2) Use the corresponding offsets to remap these center keypoints to the input image. (3) Define a central region for each bounding box and verify whether the central region contains the center keypoints. Note that the class labels of the checked center keypoints should be the same as those of the bounding box. (4) If a center keypoint is detected in the central
region, we preserve the bounding box. The score of the bounding box is 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 the central region, the bounding box is removed.

2) Multiresolution Detection Framework: The overall network architecture is shown in Fig. 3. The network starts with a backbone (e.g., ResNet [25], ResNeXt [61]) that extracts features based on the input image. We select C3-C5 feature maps from the backbone as the input to the feature pyramid network (FPN). Then the FPN outputs P3–P7 feature maps as the final prediction layers. In each prediction layer, we use a heatmap and regression to predict the keypoint. In the heatmap-based prediction method, we use three light binary heatmaps to predict the corners and center keypoints. The resolution of the heatmap is the same as that of prediction layer, therefore, we predict an additional offset for each keypoint to learn to remap the keypoint from the heatmap to the input image. In the regression-based prediction method, to decouple the top-left and bottom-right corners, we divide the ground-truth boxes into four sub-ground-truth boxes along the geometric center and select the top-left and bottom-right sub-ground-truth boxes to supervise the regression process. During the inference process, the regressed vectors act as cues to identify the closest keypoints in the corresponding heatmaps to refine keypoint locations. Finally, the predicted keypoint triplets are used to determine the final bounding boxes.

Fig. 3. Multiresolution detection framework of CenterNet. A convolutional backbone network outputs three feature maps, which are denoted as C3-C5, and input into a feature pyramid network (FPN). Then, the FPN outputs P3-P7 feature maps as the final prediction layers. In each prediction layer, we use a heatmap and regression to predict the keypoints. In the heatmap-based prediction approach, we predict three light binary heatmaps for predicting the corners and center keypoints. In the regression-based prediction approach, to decouple the top-left and bottom-right corners, we divide the ground-truth boxes into four sub-ground-truth boxes along the geometric center and select the top-left and bottom-right sub-ground-truth boxes to supervise the regression process. During the inference process, the regressed vectors act as cues to identify the closest keypoints in the corresponding heatmaps to refine keypoint locations. Finally, the predicted keypoint triplets are used to determine the final bounding boxes.

to refine the locations of the keypoints. This method successfully reduces the false positive corners introduced by heatmaps. Next, each valid pair of keypoints is used to define a bounding box. Here, valid indicates that two keypoints belong to the same class (i.e., the corresponding top-left and bottom-right sub-bounding boxes of the same class), and the x and y coordinates of the top-left point are smaller than those of the bottom-right point. Finally, we define a central region for each bounding box and verify whether the central region contains both the predicted center keypoints. If and only if there are two center keypoints detected in the central region, the bounding box is preserved, otherwise the bounding box is removed. The score of the bounding box is replaced by the average scores of the points, i.e., the scores of the top-left corner, bottom-right corner and the center keypoints.

3) Central Region Definition: The size of the central region in the bounding box affects the detection results. For example, small central regions lead to low recall rates for small bounding boxes, while large central regions lead to low precision scores for large bounding boxes. Therefore, we propose a scale-aware central region that is adaptively adjusted based on the size of the bounding boxes. Let $tl_i$ and $br_i$ denote the coordinates of the top-left corner of $i$ and $br_i$ and $br_y$ denote the coordinates of the bottom-right corner of $i$. When we define a central region $j$. Let $ctl_j$ 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 $tl_{ij}$, $tl_i$, $br_i$, $br_y$, $ctl_j$, $ctl_y$, $cbr_x$ and $cbr_y$ should satisfy the following relationships:

$$
\begin{align*}
ctl_x &= \frac{(n+1)tl_x + (n-1)br_x}{2n} \\
ctl_y &= \frac{(n+1)tl_y + (n-1)br_y}{2n} \\
cbr_x &= \frac{(n-1)tl_x + (n+1)br_x}{2n} \\
cbr_y &= \frac{(n-1)tl_y + (n+1)br_y}{2n}
\end{align*}
$$

where $n$ is odd and determines the scale of the central region $j$. In this paper, $n$ is set to 3 and 5 for bounding boxes with scales less than and greater than 150, respectively. Fig. 4 shows two central regions with $n = 3$ and $n = 5$. We can use (1) to...
determine a scale-aware central region and verify whether the central region contains the center keypoints.

C. Enriching the Center and Corner Information

The center keypoints and corners both have rigorous geometric relationships with the objects but contain limited visual patterns of the objects. We train the network in a fully supervised manner to learn these geometric relationships and the limited visual features to locate the keypoints. If more visual patterns are introduced for the center keypoints and corners, they can be detected better.

Center Pooling: The geometric centers of the objects do not always convey 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. Fig. 5(a) shows the principle of center pooling. The detailed process of center pooling is set as follows: the backbone outputs a feature map and to determine whether a pixel in the feature map is a center keypoint, we need to find the maximum value in both the boundary directions and the internal directions of the objects.

Fig. 5. (a) Center pooling determines the maximum values in the horizontal and vertical directions. (b) Corner pooling uses only the maximum values in the boundary directions. (c) Cascade corner pooling applies the maximum values in both boundary directions and the internal directions of the objects.

Cascade Corner Pooling: Corners are often outside objects and lack local appearance features. CornerNet [30] uses corner pooling to address this issue. The principle of corner pooling is shown in Fig. 5(b). Corner pooling aims to find the maximum value on the boundary and then searches inside the box at the location of the maximum boundary value to determine the internal maximum value; then, the two maximum values are added together. The cascade corner pooling allows the corners to obtain both the boundary information and the visual patterns of objects.

Center pooling and cascade corner pooling can be achieved by applying corner pooling [30] in different directions. Fig. 6(a) shows the structure of the center pooling module. To determine the maximum value in a specific direction, e.g., the horizontal direction, we only need to connect the left and right pooling in sequence. Fig. 6(b) shows the structure of the cascade top corner pooling module, in which the white rectangle denotes a $3 \times 3$ convolution followed by batch normalization.

D. Training and Inference

Training: We train CenterNet on 8 Tesla V100 (32 GB) GPUs. For the single-resolution detection framework, our baseline is CornerNet [30]. Next, we use the stacked hourglass network (Hourglass) [43] with 52 and 104 layers as the backbone; the latter has two hourglass modules, while the former has only one. All modifications to the hourglass architecture made in [30] are preserved. The network is trained from scratch when we use Hourglass as the backbone. In addition, to show that the framework can be generalized to other network architectures, we investigate another backbone named HRNet [54], [55], which maintains high-resolution representations during the feature extraction process. The resolution of the input image is 511 × 511, leading to $128 \times 128$ heatmaps. We use the data augmentation strategy presented in [30] to train a robust model. Adam [26] is used to optimize the training loss:

$$L = I_{kp}^{co} + I_{kp}^{ce} + \alpha I_{pull}^{co} + \beta I_{push}^{co} + \gamma (I_{off}^{co} + I_{off}^{ce}),$$

(2)

where $I_{kp}^{co}$ and $I_{kp}^{ce}$ denote the focal losses, which are used to train the network to detect corners and center keypoints, respectively. $I_{pull}^{co}$ is a “pull” loss for the corners, which is used to minimize the distance between the embedding vectors that belong to the same objects. $I_{push}^{co}$ is a “push” loss for the corners that is used to maximize the distance between the objects.

For the topmost, leftmost, bottommost and rightmost boundaries, look vertically downward, horizontally toward the right, vertically toward the top and horizontally toward the left, respectively.
embedding vectors that belong to different objects. \( L^{\text{cls}}_{\text{off}} \) and \( L^{\text{reg}}_{\text{off}} \) are \( \ell_1 \)-losses [21], which are used to train the network to predict the offsets of the corners and center keypoints, respectively. \( \alpha, \beta \) and \( \gamma \) denote the weights for the corresponding losses and are set to 0.1, 0.1 and 1, respectively. \( L_{\text{kp}}, L_{\text{pull}}, L_{\text{push}} \) and \( L_{\text{off}} \) are all defined in CornerNet, and the details are provided in [30]. We use a batch size of 48. The maximum number of training epochs is 100. We use a learning rate of \( 2.5 \times 10^{-4} \) for the first 88 epochs and continue training for an additional 12 epochs with a learning rate of \( 2.5 \times 10^{-5} \).

For the multiresolution detection framework, we use ResNet [25], Res2Net [18], ResNetXt [61] and Swin-Transformer [39] with the weights pretrained based on ImageNet [11] as our backbones. An FPN [35] is used to output detection layers with different scales. Single-scale and multiscale training strategies are both applied. For the single-scale training, the shorter side of each input image is 800 pixels, while for the multiscale training, the shorter side of each input image is randomly selected in the range of [480, 960]. We use the data augmentation strategy presented in [63] to train a robust model. Stochastic gradient descent (SGD) is used to optimize the training loss as follows:

\[
L_{\text{m}} = \frac{1}{2} \left( L^{\text{cls}}_{\text{cls}} + L^{\text{reg}}_{\text{cls}} \right) + \frac{1}{2} \left( L^{\text{reg}}_{\text{reg}} + L^{\text{reg}}_{\text{reg}} \right) + \hat{\alpha} \left( L^{\text{cls}}_{\text{off}} + L^{\text{reg}}_{\text{off}} \right) + \hat{\beta} \left( L^{\text{cls}}_{\text{off}} + L^{\text{reg}}_{\text{off}} \right) + \hat{\gamma} \left( L^{\text{cls}}_{\text{off}} + L^{\text{reg}}_{\text{off}} \right),
\]

where \( L^{\text{cls}}_{\text{cls}}, L^{\text{reg}}_{\text{cls}} \) denote the focal losses, which are used to train the network to classify the top-left and bottom-right sub-bounding boxes, respectively. \( L^{\text{reg}}_{\text{reg}} \) and \( L^{\text{reg}}_{\text{reg}} \) denote the GloU loss [49], which are used to train the network to regress the top-left and bottom-right sub-bounding boxes, respectively. \( \alpha, \beta \) and \( \gamma \) denote the weights of the corresponding losses and are set to 2, 0.25 and 1.0, respectively. We use a batch size of 16. The maximum number of training epochs is 24. We use a learning rate of 0.01 for the first 16 epochs, and then the learning rate is decayed by a factor of 10 after the 16th and the 22nd epoch.

**Inference:** For the single-resolution detection framework, we follow the process described in [30]. For the single-scale testing, we input the original and horizontally flipped images with the original resolutions into the network. For multiscale testing, we input the original and horizontally flipped images with resolutions of 0.6, 1, 1.2, 1.5 and 1.8. We select the 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 the top 100 bounding boxes according to their scores as the final detection results.

For multiresolution detection framework, we follow the process described in [63]. For the single-scale testing, we resize each image according to the shorter side of 800 pixels as inputs to the network, while for the multiscale testing, we resize each image according to the shorter side of [400, 600, 800, 1000, 1200, 1400], and the detection results at all scales are combined. NMS with a threshold of 0.6 is applied to remove the redundant results. Flip argumentation is used in only the multiscale evaluation.

**E. Relationship with Prior Works**

Our approach has the advantages of both bottom-up and top-down approaches. Top-down approaches can perceive the global visual contents within proposals; however, they usually suffer from low location accuracies especially for objects with peculiar shapes. Bottom-up approaches can locate objects with arbitrary geometries but often generate many incorrect bounding boxes (false positives). Our approach uses a triplet of keypoints to represent an object; thus, our model is still a bottom-up approach that can perceive visual contents within proposals and has minimal costs.

**IV. EXPERIMENTS**

**A. Dataset, Metrics and Baseline**

We evaluate our method based on the MS-COCO dataset [37]. This dataset 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 ‘train2017’ set (i.e., 110 K training images) for training and test the model based on the test-dev set. We use the ‘val2017’ set as the validation set to perform ablation studies and visualization experiments.

The MS-COCO dataset [37] uses the AP and AR metrics to characterize the detector performance. The AP represents the average precision, which is computed based on ten different IoU thresholds (i.e., \( 0.5 : 0.05 : 0.95 \)) for all categories. The AR represents the maximum recall rate, which is computed based on 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, the AP and AR can be used to evaluate the performance at different object scales, including small objects (area < \( 32^2 \)), medium objects (\( 32^2 < \text{area} < 96^2 \)) and large objects (\( \text{area} > 96^2 \)). The AP is considered the most important metric on the MS-COCO dataset.

**B. Comparisons With State-of-The-Art Detectors**

Table III shows a comparison between the proposed approach and state-of-the-art detectors based on the MS-COCO test-dev set. Compared with the baseline CornerNet [30], the proposed CenterNet shows remarkably improved performance. For example, SR-CenterNet (Hourglass-52) obtains a single-scale testing AP of 41.6%, an improvement of 3.8% over the AP of 37.8% obtained by CornerNet, and a multiscale testing AP of 43.5%, an improvement of 4.1% over the AP of 39.4% achieved by CornerNet under the same settings. When using a deeper backbone (i.e., Hourglass-104), CenterNet achieves improved AP over CornerNet [30], with improvements of 4.4% (from 40.5% to 44.9%) and 4.9% (from 42.1% to 47.0%) under the single-scale and multiscale testing conditions, respectively. We also report the detection results of MR-CenterNet, which obtain significant improvements on the basis of SR-CenterNet, MR-CenterNet, which uses Res2Net-101 as a backbone, obtains APs of 51.5% for single-scale testing and 53.7% for multiscale testing. The
current Transformer-based top-down detectors [39] achieves state-of-the-art accuracy, we explored the application of Transformer-based backbones in bottom-up approaches. The CenterNet with a Transformer-based backbone achieves APs of 53.2% for single-scale testing and 57.1% for multiscale testing, outperforming all other bottom-up approaches to the best of our knowledge.

The largest improvement is observed for small objects. For instance, SR-CenterNet (Hourglass-104) improves the AP for small objects by 4.7% (from 42.7% to 47.4%) and 3.5% (from 53.9% to 57.4%) for medium and large bounding boxes, respectively. Fig. 7(c) and (d) show qualitative comparisons of the reduction in the number of incorrect medium and large bounding boxes. Notably, the AR is also significantly improved compared with the AR scores of the baselines, with the best performance with the AR scores of the baselines, with the best performance.
Fig. 7. (a) and (b) show that the number of incorrect small bounding boxes is significantly reduced by modeling the center information. (c) and (d) show that the center information also helps to reduce the number of incorrect medium and large bounding boxes. (e) shows the center keypoint detection results without and with center pooling. (f) shows the corner detection results with corner pooling and cascade corner pooling. The blue boxes denote the ground truth. The red boxes and dots denote the predicted bounding boxes and keypoints, respectively.

Fig. 8. Some qualitative detection results based on the MS-COCO validation dataset. Only detections with scores higher than 0.5 are shown.

achieved under the multiscale testing condition. This is because our approach removes many incorrect bounding boxes, which is equivalent to improving the confidence of bounding boxes with accurate locations but relatively low scores.

The performance of CenterNet is also competitive when compared with that of top-down approaches, e.g., the single-scale testing AP of SR-CenterNet (Hourglass-52) is comparable to that of the top-down approach RefineDet [65] (41.6% versus 41.8%), and the single-scale testing AP of MR-CenterNet (Res2Net-101) is comparable that of GFLV2 [32] (53.7% versus 53.3%). The multi-scale testing AP of 57.1% achieved by MR-CenterNet (Swin-L) closely matches the state-of-the-art AP of 58.7% achieved by Swin Transformer [39], a top-down approach. We present some qualitative detection results in Fig. 8.

C. Multiresolution Detection Improves Precision

As shown in Table III, the proposed MR-CenterNet improves the object detection accuracy when compared with existing approaches. For instance, at the same network depth (Hourglass-52 versus ResNet-50), MR-CenterNet improves the AP of objects by 4.8%. Moreover, due to the strong generalizability of the MR-CenterNet framework, we can apply stronger backbones for CenterNet.

We design two comparative experiments to confirm the contribution of the multiresolution detection process to the performance of MR-CenterNet. The first experiment is a control experiment, so we use the default network. For the second experiment, we augment the network with two up-convolutional networks and skip connections from the lower layers to the output to obtain higher-resolution outputs. The resolution of the output layer is 1/8 times that of the input image. Therefore, all the objects are detected in a single-resolution detection layer. Diagrams of the two network structures are shown in Fig. 9. Table IV reports the detection results of the two experiments based on the MS-COCO validation dataset. MR-CenterNet with multiresolution detection layers achieves higher accuracy than the other models. The multiresolution detection structure provides richer receptive fields for detecting objects with different scales, which helps to improve the detection accuracy.

D. Real-Time CenterNet

We also design a real-time version of CenterNet, called CenterNet-RT. CenterNet-RT is based on the multiresolution
Our work has three main contributions, including central region exploration, center pooling and cascade corner pooling. To analyze the contribution of each individual component, we performed ablation studies. The baseline is CornerNet511-52 [30]. We add the three components to the baseline one by one and follow the default parameter setting detailed in Section IV-A. The results are shown in Table VII.

Central Region Exploration: To understand the importance of 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 the bounding boxes. For center keypoint detection, we use a good trade-off between accuracy and speed. In our conference version, CenterNet (SR-CenterNet) performs slowly, with an inference speed of less than 7 FPS. In this paper, we equip CenterNet with a pyramid structure and detect objects in multiresolution feature layers, which significantly improves the speed and accuracy. With ResNet-50, MR-CenterNet achieves an AP of 45.7% at 14.5 FPS. Furthermore, we propose CenterNet-RT on the basis of MR-CenterNet, which achieves an AP of 43.2% at 30.5 FPS. This accuracy is competitive with that of SR-CenterNet (HG-104), but the inference speed is ~ 6 times faster.

E. Incorrect Bounding Box Reduction

The AP [37] metric reflects the number of high quality object bounding boxes (usually IoU \( \geq 0.5 \)) that can be predicted by a network; however, this metric does not directly reflect how many incorrect object bounding boxes (usually IoU \( \leq 0.5 \)) are generated by the network. The AF rate is a suitable metric, that reflects the proportion of incorrect bounding boxes. Table VI shows the AF rates for CornerNet and CenterNet. CornerNet generates many incorrect bounding boxes even at an IoU = 0.05 threshold, i.e., CornerNet511-52 and CornerNet511-104 obtain AF scores of 35.2% and 32.7%, respectively. On the other hand, CornerNet generates more small incorrect bounding boxes than medium and large incorrect bounding boxes, with AF scores of 62.5% for CornerNet511-52 and 60.3% for CornerNet511-104. Our CenterNet decreases the AF rates based on all criteria by exploring the central regions. For instance, CenterNet511-52 and CornerNet511-104 obtain AF scores of 9.5% for CenterNet511-52 and 6.8% for CornerNet511-104. This is also the reason why the AP improvement is more prominent for small objects than for medium and large objects.

F. Ablation Study

Our work has three main contributions, including central region exploration, center pooling and cascade corner pooling. To analyze the contribution of each individual component, we performed ablation studies. The baseline is CornerNet511-52 [30]. We add the three components to the baseline one by one and follow the default parameter setting detailed in Section IV-A. The results are shown in Table VII.

Central Region Exploration: To understand the importance of 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 the bounding boxes. For center keypoint detection, we use a good trade-off between accuracy and speed. In our conference version, CenterNet (SR-CenterNet) performs slowly, with an inference speed of less than 7 FPS. In this paper, we equip CenterNet with a pyramid structure and detect objects in multiresolution feature layers, which significantly improves the speed and accuracy. With ResNet-50, MR-CenterNet achieves an AP of 45.7% at 14.5 FPS. Furthermore, we propose CenterNet-RT on the basis of MR-CenterNet, which achieves an AP of 43.2% at 30.5 FPS. This accuracy is competitive with that of SR-CenterNet (HG-104), but the inference speed is ~ 6 times faster.

E. Incorrect Bounding Box Reduction

The AF [37] metric reflects the number of high quality object bounding boxes (usually IoU \( \geq 0.5 \)) that can be predicted by a network; however, this metric does not directly reflect how many incorrect object bounding boxes (usually IoU \( \leq 0.5 \)) are generated by the network. The AF rate is a suitable metric, that reflects the proportion of incorrect bounding boxes. Table VI shows the AF rates for CornerNet and CenterNet. CornerNet generates many incorrect bounding boxes even at an IoU = 0.05 threshold, i.e., CornerNet511-52 and CornerNet511-104 obtain AF scores of 35.2% and 32.7%, respectively. On the other hand, CornerNet generates more small incorrect bounding boxes than medium and large incorrect bounding boxes, with AF scores of 62.5% for CornerNet511-52 and 60.3% for CornerNet511-104. Our CenterNet decreases the AF rates based on all criteria by exploring the central regions. For instance, CenterNet511-52 and CornerNet511-104 both decrease the AF score by 4.5%. In addition, the AF rates for small bounding boxes decrease the most, with AF scores of 9.5% for CornerNet511-52 and 9.6% for CenterNet511-104. This is also the reason why the AP improvement is more prominent for small objects than for medium and large objects.

F. Ablation Study

Our work has three main contributions, including central region exploration, center pooling and cascade corner pooling. To analyze the contribution of each individual component, we performed ablation studies. The baseline is CornerNet511-52 [30]. We add the three components to the baseline one by one and follow the default parameter setting detailed in Section IV-A. The results are shown in Table VII.

Central Region Exploration: To understand the importance of 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 the bounding boxes. For center keypoint detection, we use
only conventional convolutions. As presented in the third row in Table VII, the AP is improved by 2.3% (from 37.6% to 39.9%) when central region exploration is added. Moreover, we find that the improvement for 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 the number of small incorrect bounding boxes is larger, and these bounding boxes usually do not contain the center keypoints of the objects and are thus, more likely to benefit from filtering by the center keypoints.

Center Pooling: To demonstrate the effectiveness of the proposed center pooling, we next add the center pooling module to the network (see CTP in the table). The fourth row in Table VII shows the addition of the center pooling improves the AP by 0.9% (from 39.9% to 40.8%). Notably, with the addition of the center pooling, the AP for large objects is improved by 1.4% (from 52.2% to 53.6%), which is a much large improvement that achieved by using conventional convolutions (i.e., 1.4% versus 0.1%). This result demonstrates that our center pooling is effective for detecting the center keypoints of objects, especially for large objects. Our explanation is that center pooling can be used to extract richer internal visual patterns, and larger objects contain more accessible internal visual patterns than smaller objects. Fig. 7(e) shows the results of detecting center keypoints without and with center pooling. The model with the conventional convolution fails to locate the center keypoint for the cow, while the model with center pooling successfully locates the center keypoint.

Cascade Corner Pooling: We replace the corner pooling [30] with the cascade corner pooling to detect the corners (see CCP in the table). The second row in Table VII shows the results on the basis of CornerNet511-52. We find that the addition of the cascade corner pooling improves the AP by 0.7% (from 37.6% to 38.3%). The last row shows the results on the basis of CenterNet511-52, demonstrating that the AP is improved by 0.5% (from 40.8% to 41.3%). The results in the second row show that there is almost no change in the AP for large objects (i.e., 52.2% versus 52.2%), but the AR is improved by 1.8% (from 74.0% to 75.8%). This result suggests that the cascade corner pooling helps to obtain more internal visual patterns within the bounding box. To understand the importance of the center keypoints, we replace the predicted center keypoints with the ground-truth values and evaluate the model performance based on the MS-COCO validation dataset. Table VIII 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.

G. 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 misses the visual patterns within the bounding box. To understand the importance of the center keypoints, we replace the predicted center keypoints with the ground-truth values and evaluate the model performance based on the MS-COCO validation dataset. Table VIII 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.

V. Conclusion

In this paper, we propose CenterNet, a new bottom-up object detection approach that detects objects using a triplet of keypoints, including one center keypoint and two corners. Our approach addresses the problem that traditional bottom-up approaches lack additional investigations into the cropped regions by exploring the visual patterns within each proposed region with minimal costs. In addition, we extend CenterNet based on a framework with a pyramid structure to improve the multiscale object detection performance. The experimental results show that CenterNet outperforms all existing bottom-up approaches by a large margin and is competitive when compared with the state-of-the-art top-down approaches, especially in terms of the recall rate. We also design some real-time CenterNet models, which achieve a good trade-off between accuracy and speed.

Importantly, we prove that bottom-up approaches are more flexible than top-down approaches in locating objects with arbitrary geometries and that additional investigations into each

---

**Table VII**

| Component         | AP   | AP50  | AP75  | APs   | APM   | APL   | AR1   | AR10  | AR100  | ARs   | ARM   | ARL   |
|-------------------|------|-------|-------|-------|-------|-------|-------|-------|--------|-------|-------|-------|
| CRE               | 37.6 | 53.3  | 40.0  | 18.5  | 39.6  | 52.2  | 33.7  | 52.2  | 56.7   | 37.2  | 60.0  | 74.0  |
| CTP               | ✓    | 38.3  | 54.2  | 40.5  | 18.6  | 40.5  | 52.2  | 34.0  | 53.0   | 57.9  | 36.6  | 60.8  | 75.8  |
| CCP               | ✓ ✓  | 39.9  | 57.7  | 42.3  | 23.1  | 42.3  | 52.3  | 33.8  | 54.2   | 58.5  | 38.7  | 62.4  | 74.4  |
| CCP               | ✓ ✓  | 40.8  | 58.6  | 43.6  | 23.6  | 43.6  | 53.6  | 33.9  | 54.5   | 59.0  | 39.0  | 63.2  | 74.7  |
| CCP               | ✓ ✓  | 41.3  | 59.2  | 43.9  | 23.6  | 43.8  | 55.8  | 34.5  | 55.0   | 59.2  | 39.1  | 63.5  | 75.1  |

CRE denotes central region exploration, CTP denotes center pooling, and CCP denotes cascade corner pooling.

**Table VIII**

| Method                  | Backbone | AP   | AP50  | AP75  | APs   | APM   | AP L  | AR1   | AR10  | AR100  | ARs   | ARM   | ARL   |
|-------------------------|----------|------|-------|-------|-------|-------|-------|-------|-------|--------|-------|-------|-------|
| SR-CenterNet w/o GT     | HC-52    | 41.3 | 57.2  | 43.9  | 23.6  | 43.8  | 55.8  |       |       |        |       |       |       |
| SR-CenterNet w/ GT      | HC-52    | 56.5 | 78.3  | 61.4  | 39.1  | 60.3  | 70.3  |       |       |        |       |       |       |
| SR-CenterNet w/o GT     | HC-104   | 44.8 | 62.4  | 48.2  | 25.9  | 48.9  | 58.8  |       |       |        |       |       |       |
| SR-CenterNet w/ GT      | HC-104   | 58.1 | 78.4  | 63.9  | 40.4  | 63.9  | 72.1  |       |       |        |       |       |       |

We replace the predicted center keypoints with the ground-truth values, and the results suggest that there is still room for improvement in detecting the center keypoints.
proposed region are necessary to improve model precision. We hope that CenterNet will attract more attention and promote further exploration of bottom-up approaches.

REFERENCES

[1] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, “Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2874–2883.

[2] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, “Soft-NMS—improving object detection with one line of code,” in Proc. IEEE Int. Conf. on Comput. Vis., 2017, pp. 556–565.

[3] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 6154–6162.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 729–739.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A.Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 213–229.

[6] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, “Blendmask: Top-down meets bottom-up for instance segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 8573–8581.

[7] Y. Chen, C. Han, N. Wang, and Z. Zhang, “Revisiting feature alignment for one-stage object detection,” 2019, arXiv: 1908.01570.

[8] Y. Chen, Z. Zhang, Y. Cao, L. Wang, S. Lin, and H. Hu, “Reppoints V2: Verification meets regression for object detection,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 5621–5631.

[9] J. Dai, Y. Li, K. He, and J. Sun, “R-FCN: Object detection via region-based fully convolutional networks,” in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 379–387.

[10] J. Dai et al., “Deformable convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 764–773.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 248–255.

[12] Z. Dong, G. Li, Y. Liao, F. Wang, P. Ren, and C. Qian, “CenntipetalNet: Pursuing high-quality keypoint pairs for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 10519–10528.

[13] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, “CenterNet: Keypoint triplets for object detection,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 6569–6578.

[14] K. Duan, L. Xie, H. Qi, S. Bai, Q. Huang, and Q. Tian, “Corner proposal network for anchor-free, two-stage object detection,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 399–416.

[15] K. Duan, L. Xie, H. Qi, S. Bai, Q. Huang, and Q. Tian, “Location-sensitive visual recognition with cross-IoU loss,” 2021, arXiv: 2104.04899.

[16] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[17] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, “DSSD: Deconvolutional single shot detector,” 2017, arXiv: 1701.06659.

[18] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. S. Torr, “Res2Net: A new multi-scale backbone architecture,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 2, pp. 652–662, Feb. 2021.

[19] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, “YoloX: Exceeding Yolo series in 2021,” 2021, arXiv:2107.08430.

[20] S. Gidaris and N. Komodakis, “Object detection via a multi-region and semantic segmentation-aware CNN model,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 1134–1142.

[21] R. Girshick, “Fast R-CNN,” in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1440–1448.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 580–587.

[23] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 2, pp. 386–397, Feb. 2020.

[24] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[25] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[26] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.
[54] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep-high-resolution representation learning for human pose estimation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.

[55] K. Sun et al., “High-resolution representations for labeling pixels and regions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, vol. 36, pp. 588–597.

[56] Z. Tian, C. Shen, H. Chen, and T. He, “FCOS: A simple and strong anchor-free object detector,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1922–1933, Apr. 2022.

[57] L. Tychsen-Smith and L. Petersson, “DeNet: Scalable real-time object detection with directed sparse sampling,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 428–436.

[58] L. Tychsen-Smith and L. Petersson, “Improving object localization with fitness NMS and bounded IoU loss,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6877–6885.

[59] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “Scaled-YOLOv4: Scaling cross stage partial network,” 2020, arXiv: 2011.08036.

[60] B. Xiao, H. Wu, and Y. Wei, “Simple baselines for human pose estimation and tracking,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 466–481.

[61] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1492–1500.

[62] B. Yang, J. Yan, Z. Lei, and S. Z. Li, “Craft objects from images,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 6043–6051.

[63] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Liu, “RepPoints: Point set representation for object detection,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9657–9666.

[64] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, “Bridge the gap between anchor-based and anchor-free detection via adaptive training sample selection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9759–9768.

[65] Z. Tian, C. Shen, H. Chen, and T. He, “FCOS: A simple and strong anchor-free object detector,” in *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018, pp. 428–436.

[66] X. Zhang, F. Wan, C. Liu, R. Ji, and Q. Ye, “Free anchor: Learning to match anchors for visual object detection,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 147–155.

[67] Q. Zhao et al., “M2Det: A single-shot object detector based on multi-level feature pyramid network,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 9259–9266.

[68] X. Zhou, D. Wang, and P. Krähenbühl, “Objects as points,” 2019, arXiv: 1904.07830.

[69] X. Zhou, J. Zhao, and P. Krähenbühl, “Bottom-up object detection by grouping extreme and center points,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 850–859.

[70] C. Zhu, F. Chen, Z. Shen, and M. Savvides, “Soft anchor-point object detection,” 2019, arXiv: 1911.12448.

[71] C. Zhu, Y. He, and M. Savvides, “Feature selective anchor-free module for single-shot object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 840–849.

[72] Z. Zou, Z. Shi, Y. Guo, and J. Ye, “Object detection in 20 years: A survey,” 2019, arXiv: 1905.05055.

Kaiwen Duan received the MS degree in pattern recognition and intelligent system from Shanghai University, Shanghai, China, in 2017, and the PhD degree in computer science from the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China, in 2022. He is currently a researcher with Baidu Inc. His current research interests include computer version, machine learning, and automatic drive.

Qi Tian (Fellow, IEEE) received the PhD degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC) in 2002. He is currently the chief scientist in artificial intelligence with Huawei Cloud and AI. He was the chief scientist in computer vision with Huawei Noah’s Ark Laboratory from 2018-2020. Before he joined Huawei, he was a full professor with the Department of Computer Science, University of Texas at San Antonio (UTSA), (2002-2019). He was listed in the Top 10 of the 2016 Most Influential Scholars in Multimedia by Aminer.org. He is an Academician of International Eurasian Academy of Sciences (IEAS), 2021. He is the recipient of the 2017 UTSA president distinguished faculty Research Achievement Award, 2016 UTSA Innovation Award in the first category, 2014 Research Achievement Awards from College of Science, UTSA, and 2010 Google Faculty Research Award. He has served as founding member of ICMR, ACM MM (2009-2012), and International Steering Committee Member for ACM MIR (2006-2010), ACM ICIMCS 2013, ICME 2006 and 2009, PCM 2012, and IEEE International Symposium on Multimedia 2011, chair for ACM Multimedia 2015. He is the associate editor of *IEEE Transactions on Multimedia*, *IEEE Transactions on Circuits and Systems for Video Technology*, *ACM Transactions on Multimedia Computing, Communications, and Applications*, *MM Science Journal*, and *MVA*.

Song Bai is a computer vision lead with ByteDance/TikTok Singapore, and also holds an appointment as an adjunct Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore. Before that, he was a research fellow with the University of Oxford working with Prof. Philip H.S. Torr. He serves as an associate editor with Pattern Recognition and guest editor of *IEEE Transactions on Pattern Analysis and Machine Intelligence*. He has served as area chair/SPC of CVPR 2023 and multiple conferences in top-tier computer vision and multimedia field. He has authored or coauthored more than 400 academic papers in prestigious international journals and top-level international conferences. He has been the associate editor of *IEEE Transactions on Circuits and Systems for Video Technology* and *Acta Automatica Sinica*, and the reviewer of various international journals including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Multimedia*, etc. He has served as general chair, program chair, area chair and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICMR, PCM, BigMM, PSIVT, etc.

Lingxi Xie received the BE and PhD degrees in engineering from Tsinghua University, in 2010 and 2015, respectively. He is currently a senior researcher with Huawei Inc. He also served as a postdoctoral researcher with the CVIL Lab from 2015 to 2019 having moved from the University of California, Los Angeles to the Johns Hopkins University. His research interests include computer vision, particular, and application of deep learning models. His research covers image classification, object detection, semantic segmentation, and other vision tasks. He is also facilitating the application of automated machine learning to the above research fields. He has published more than 90 papers in top-tier international conferences and journals.

Honggang Qi (Member, IEEE) received the MS degree in computer science from Northeast University, Shenyang, China, in 2002, and the PhD degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2008. He is currently a professor with the School of Computer Science and Technology, University of Chinese Academy of Sciences. His current research interests include computer vision, video coding, and very large scale integration design.

Qingming Huang (Fellow, IEEE) received the bachelor’s degree in computer science and the PhD degree in computer engineering from the Harbin Institute of Technology, China, in 1988 and 1994, respectively. He is currently a chair professor with the University of Chinese Academy of Sciences, China, and an adjunct research professor with the Institute of Computing Technology, Chinese Academy of Sciences, China. His research interests include multimedia computing, image processing, computer vision, and pattern recognition. He has authored or coauthored more than 400 academic papers in prestigious international journals and top-level international conferences. He has been the associate editor of *IEEE Transactions on Circuits and Systems for Video Technology* and *Acta Automatica Sinica*, and the reviewer of various international journals including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Multimedia*, etc. He has served as general chair, program chair, area chair and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICMR, PCM, BigMM, PSIVT, etc.