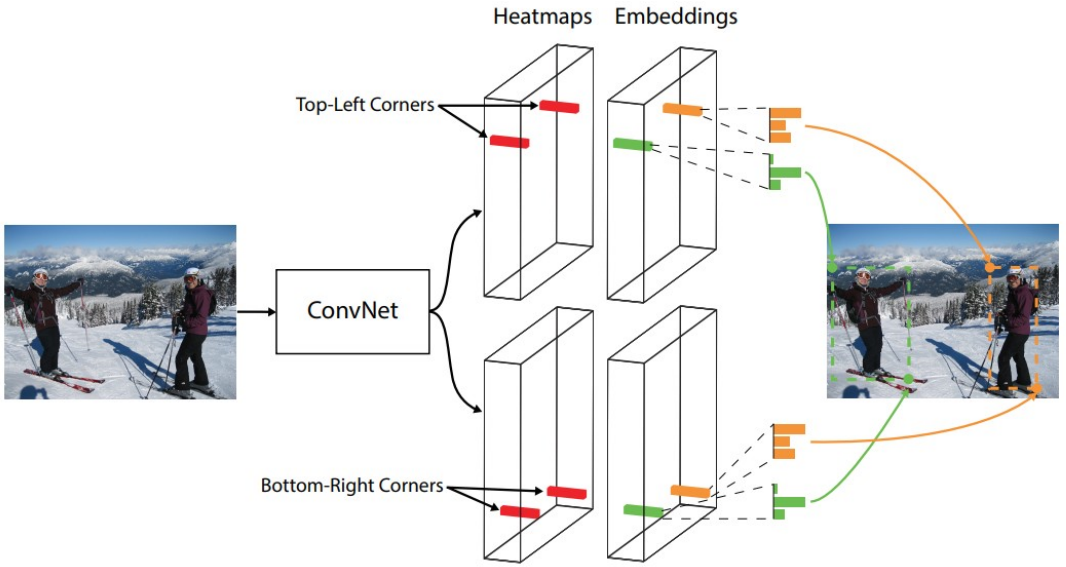
CenterNet: Keypoint Triplets for Object Detection

Seho Kim

https://arxiv.org/abs/1904.08189v3

Introduction



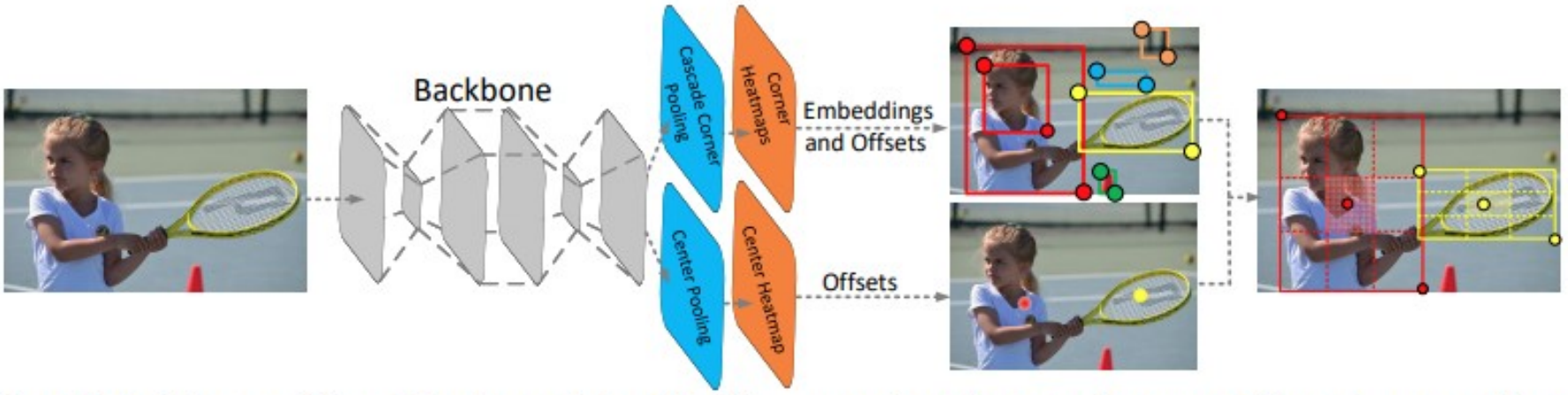


Figure 2: Architecture of CenterNet. A convolutional backbone network applies cascade corner pooling and center pooling to output two corner heatmaps and a center keypoint heatmap, respectively. Similar to CornerNet, a pair of detected corners and the similar embeddings are used to detect a potential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

Approach

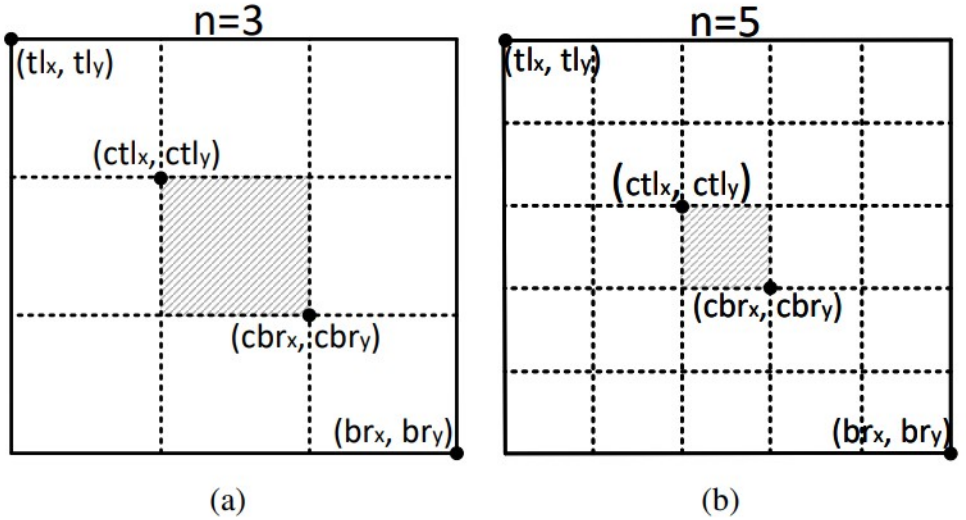
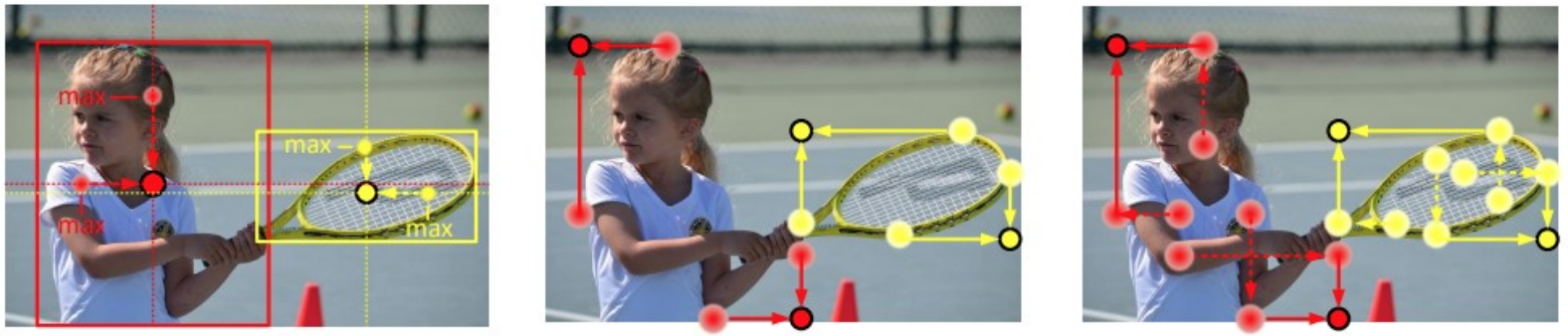


Figure 3: (a) The central region when $n = 3$. (b) The central region when $n = 5$. The solid rectangles denote the predicted bounding boxes and the shaded regions denote the central regions.

- Select top-k center keypoints according to their scores
- Use the corresponding offsets to remap these center keypoints to the input image
- Define a central region for each bounding box and check if the central region contains center keypoints

Approach

- Center pooling and Cascade corner pooling



(a) (b) (c)

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.

Approach

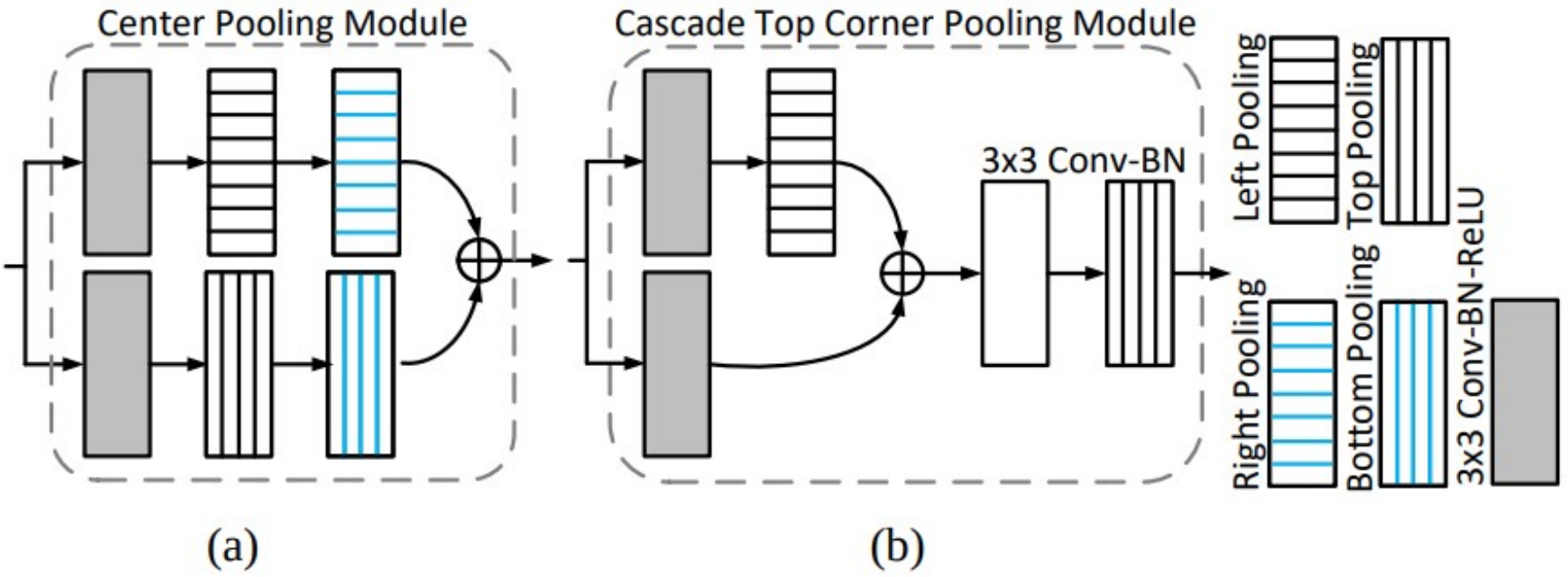


Figure 5: The structures of the center pooling module (a) and the cascade top corner pooling module (b). We achieve center pooling and the cascade corner pooling by combining the corner pooling at different directions.

Approach

- Training and Inference
  - Input image size: 511x511
  - Heatmap size: 128x128
  - Adam optimizer

$$\mathrm{L} = \mathrm{L}_{\mathrm{det}}^{\mathrm{co}} + \mathrm{L}_{\mathrm{det}}^{\mathrm{ce}} + \alpha \mathrm{L}_{\mathrm{pull}}^{\mathrm{co}} + \beta \mathrm{L}_{\mathrm{push}}^{\mathrm{co}} + \gamma \left(\mathrm{L}_{\mathrm{off}}^{\mathrm{co}} + \mathrm{L}_{\mathrm{off}}^{\mathrm{ce}} \right), \quad (2)$$

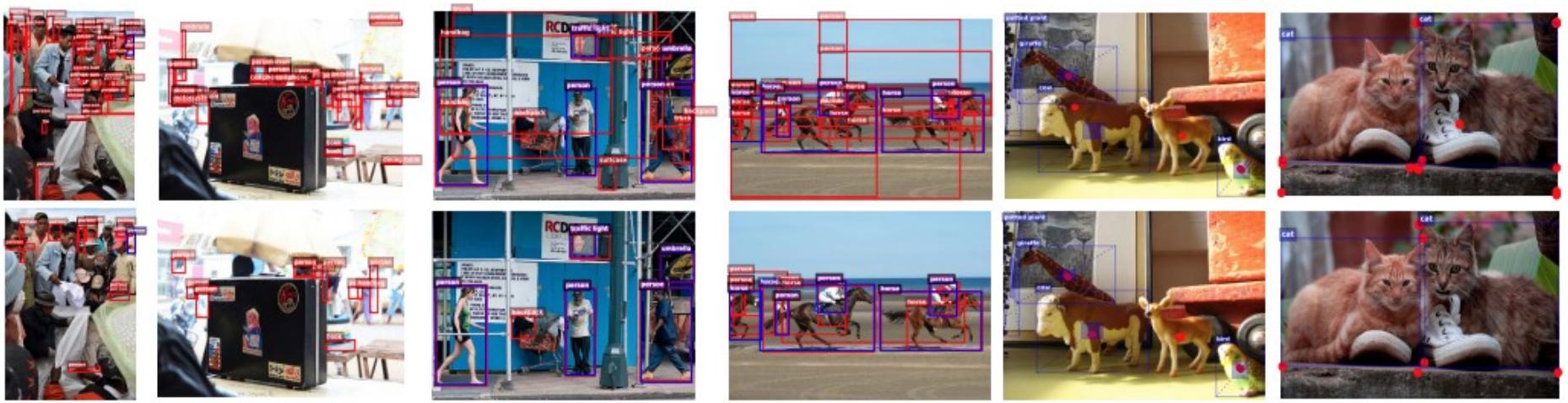
  - Focal losses: detect corners and center keypoints
  - Pull loss for corners: minimize the distance of the embedding vectors that belongs to the same objects
  - Push loss for corners: maximize the distance of the embedding vectors that belongs to different objects
  - L1 losses: train the network to predict the offsets of corners and center keypoints

Experiments

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.

Experiments



(a) (b) (c) (d) (e) (f)

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.

Experiments

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.

- Inference Speed
  - CornerNet511-104 300ms per image
  - CenterNet511-104 340ms
  - CenterNet511-52 270ms

Experiments

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.