D2Det: Towards High Quality Object Detection and Instance Segmentation

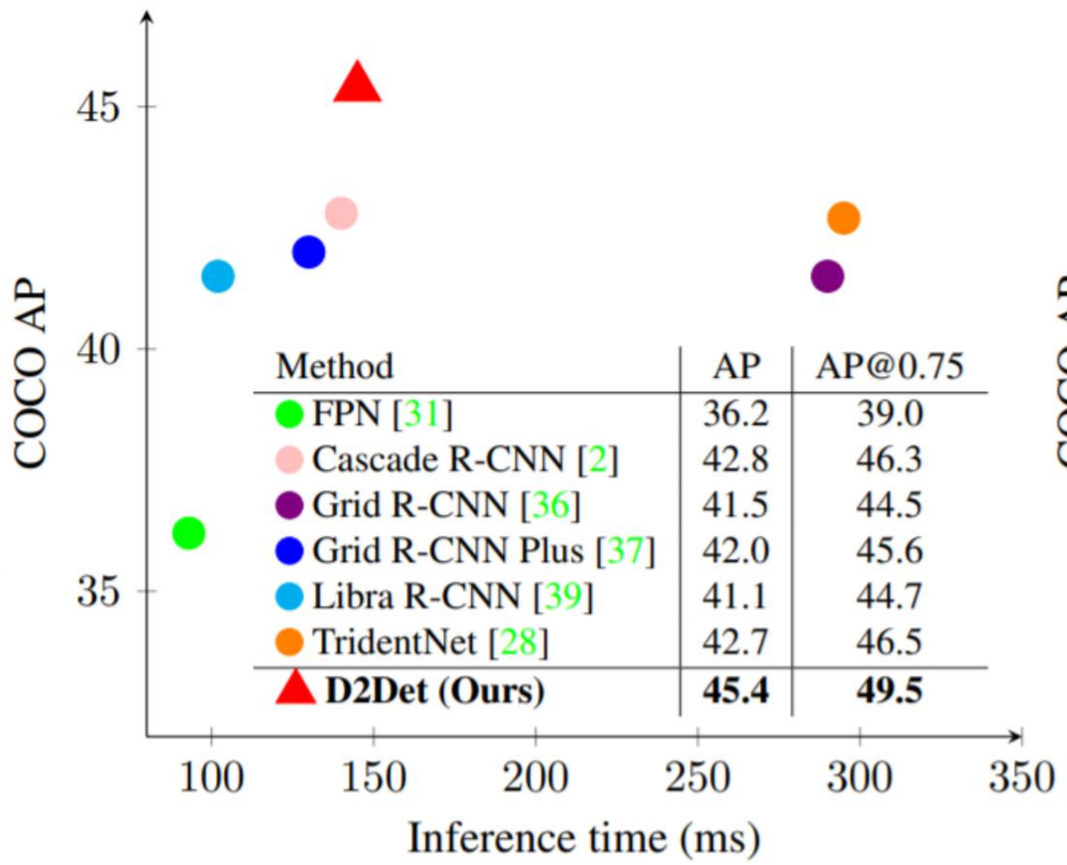
Jiale Cao*, Hisham Cholakkal*, Rao Muhammad Anwer,
Fahad Shahbaz Khan, Yanwei Pang, Ling Shao

Tianjin University, Inception Institute of Artificial Intelligence (IIAI), UAE

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

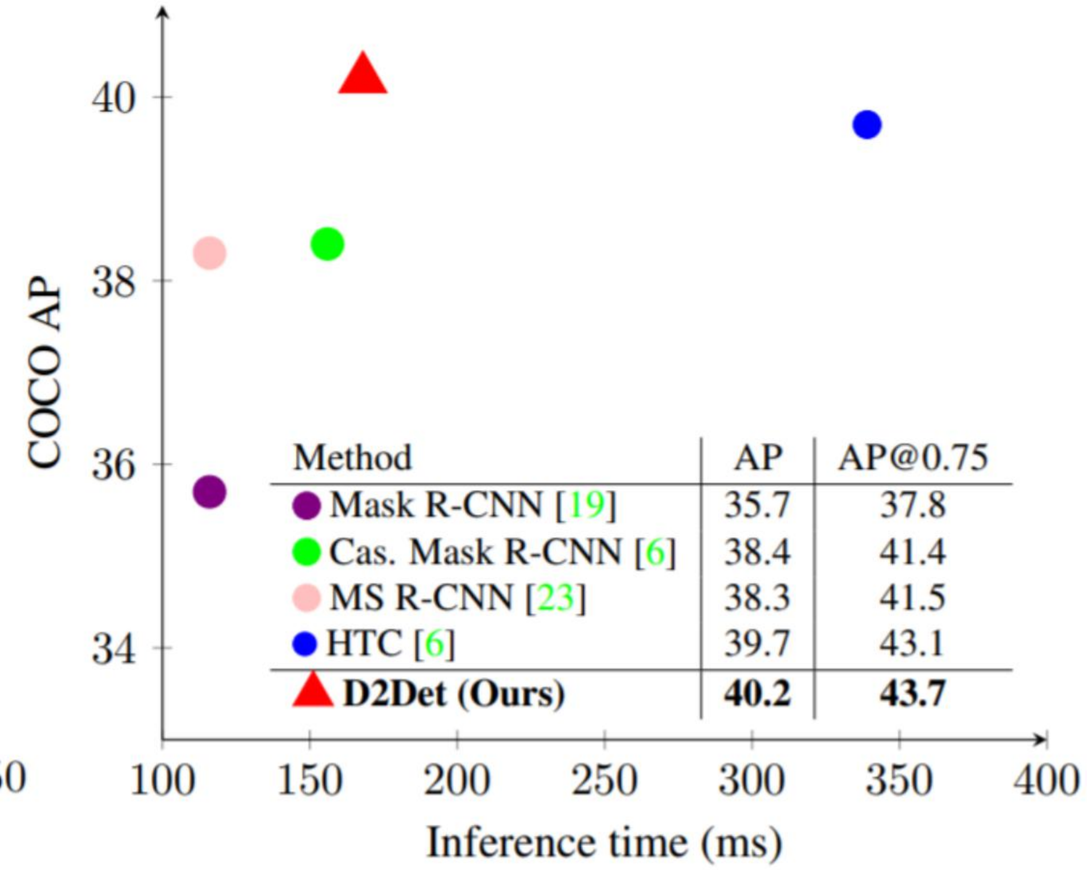
1 Introduction

- This paper proposes a novel two-stage detector (D2Det) for **high-quality** object detection and instance segmentation.



Method	AP	AP@0.75
FPN [31]	36.2	39.0
Cascade R-CNN [2]	42.8	46.3
Grid R-CNN [36]	41.5	44.5
Grid R-CNN Plus [37]	42.0	45.6
Libra R-CNN [39]	41.1	44.7
TridentNet [28]	42.7	46.5
D2Det (Ours)	**45.4**	**49.5**

(a) Object detection

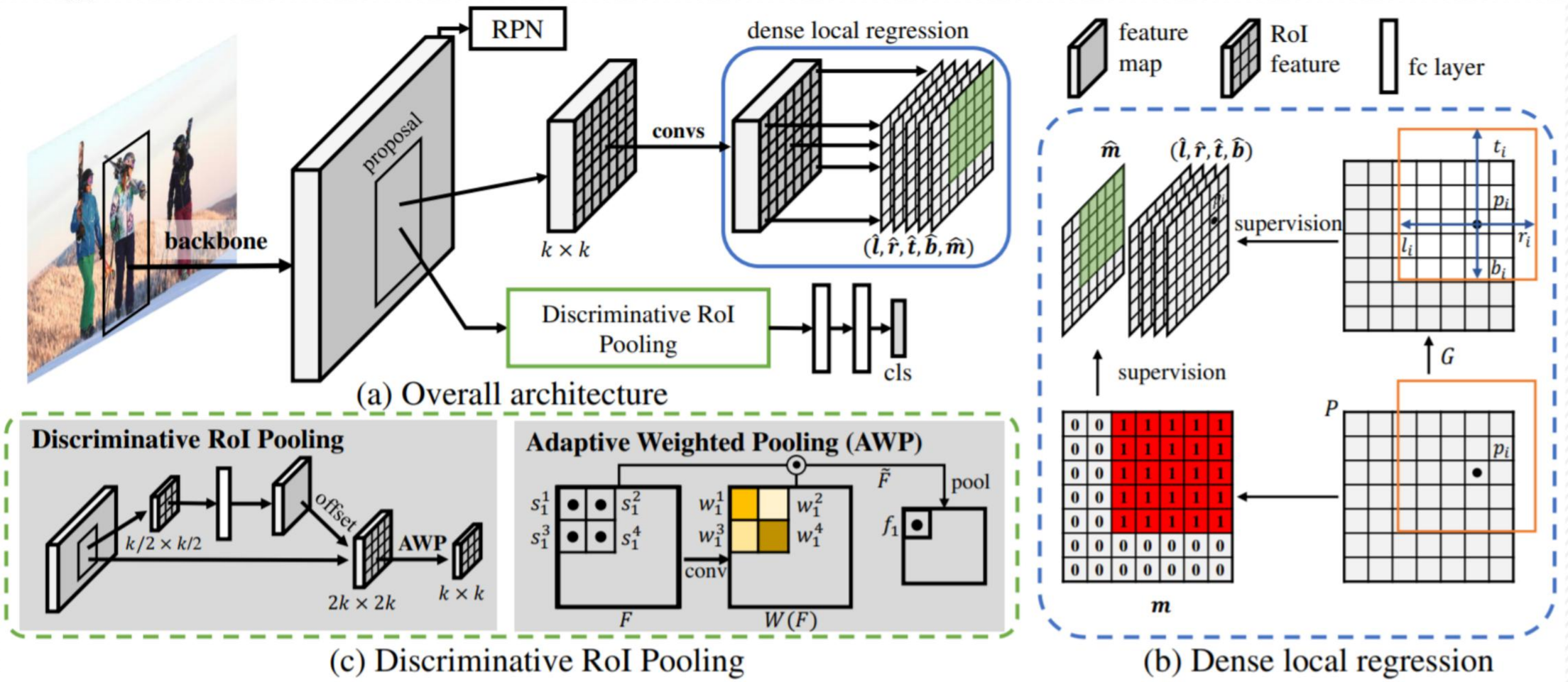


Method	AP	AP@0.75
Mask R-CNN [19]	35.7	37.8
Cas. Mask R-CNN [6]	38.4	41.4
MS R-CNN [23]	38.3	41.5
HTC [6]	39.7	43.1
D2Det (Ours)	**40.2**	**43.7**

(b) Instance segmentation

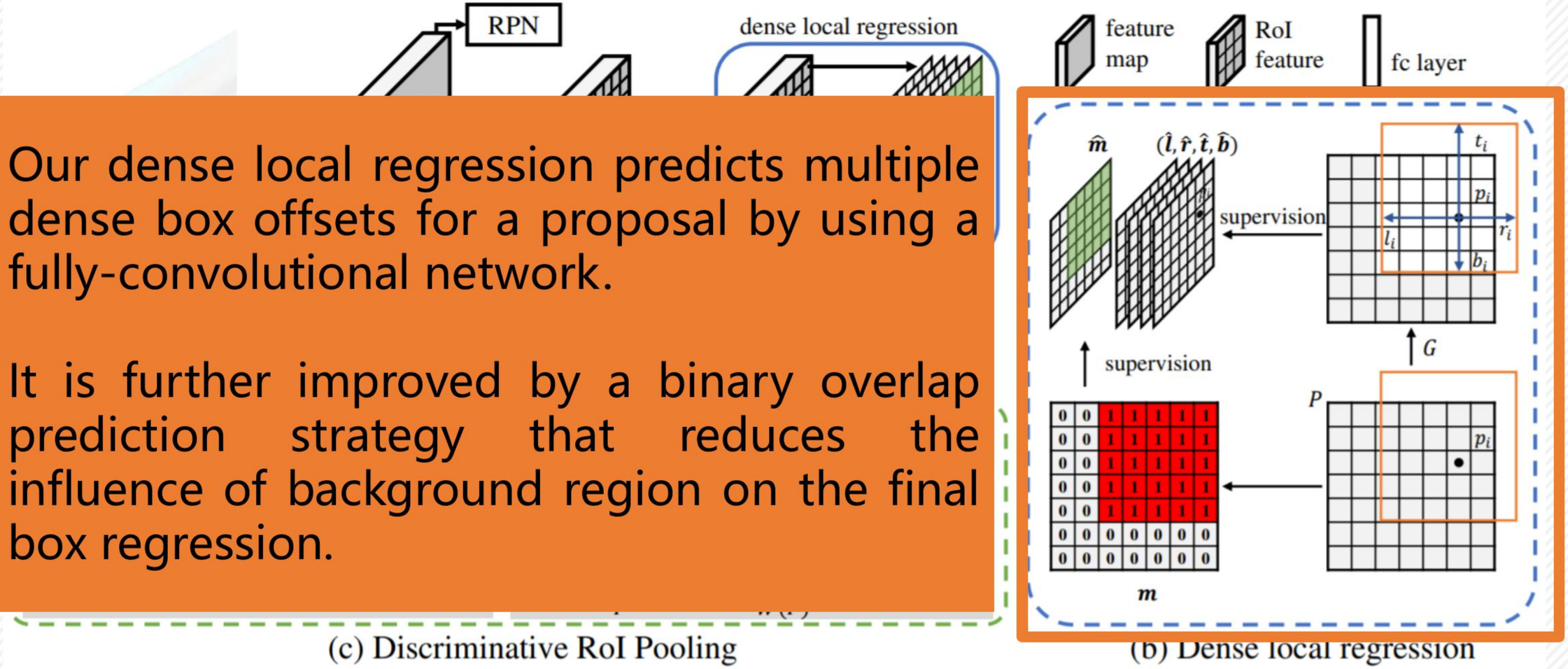
1 Our method

- D2Det re-designs both regression and classification branches of the traditional two-stage R-CNN detectors by **dense local regression** and **discriminative classification**.



1 Our method

- D2Det redesigns both regression and classification branches of the traditional two-stage R-CNN detectors by dense local regression and discriminative classification.

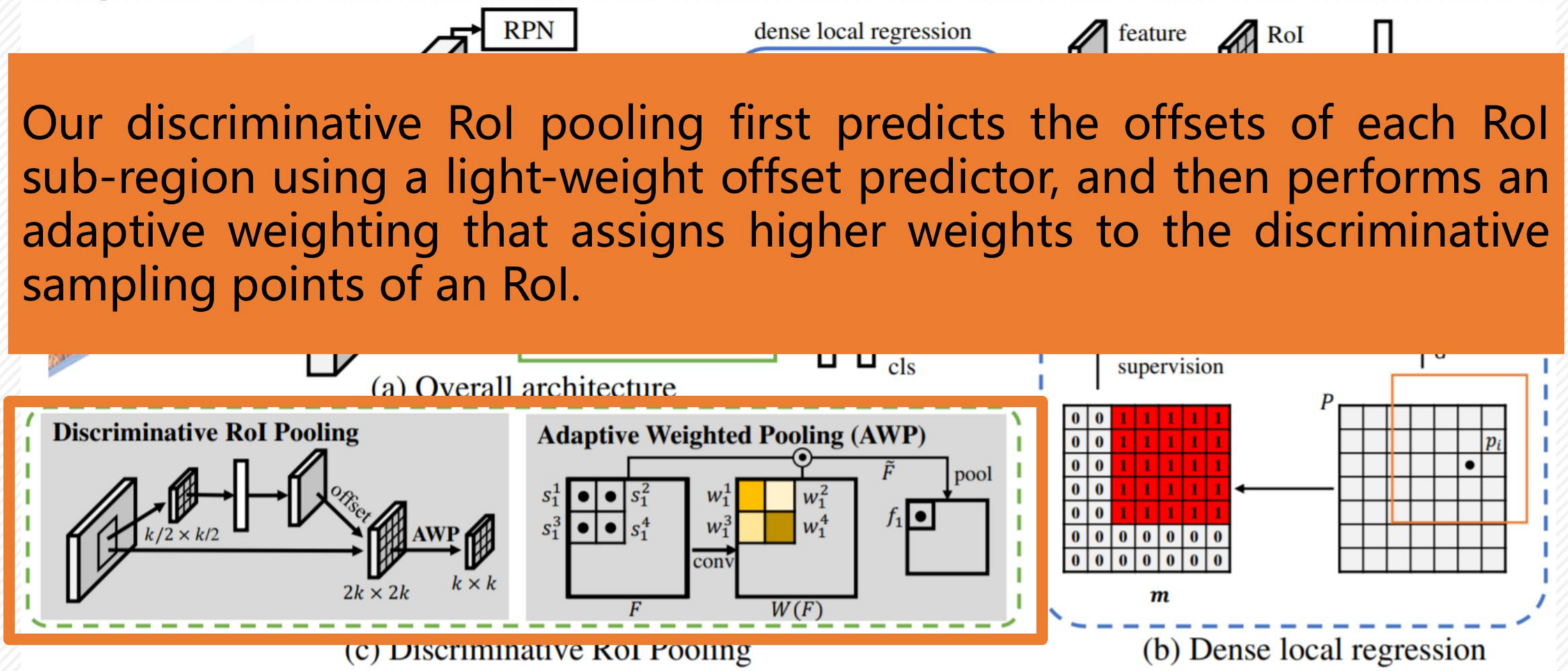


- Our dense local regression predicts multiple dense box offsets for a proposal by using a fully-convolutional network.
- It is further improved by a binary overlap prediction strategy that reduces the influence of background region on the final box regression.

1 Our method

- D2Det redesigns both regression and classification branches of the traditional two-stage R-CNN detectors by dense local regression and discriminative classification.

Our discriminative RoI pooling first predicts the offsets of each RoI sub-region using a light-weight offset predictor, and then performs an adaptive weighting that assigns higher weights to the discriminative sampling points of an RoI.



1 Our method

- D2Det redesigns both regression and classification branches of the traditional two-stage R-CNN detectors by dense local regression and discriminative classification.

- Our method is **easily** extended to **instance segmentation** by modifying the dense local regression branch.

  - Instead of assuming all regions inside the ground-truth bounding box belong to the object, the **ground-truth mask** is used to label local features.

  - we utilize **two deconvolutional layers** that increase the output spatial resolution by four times (28x28) and **two fully-connected layers** for efficient mask scoring.

2 Experiments

Methods	Backbone	Input Size	AP	AP@0.5	AP@0.75	AP_s	AP_m	AP_l
Single-Stage Methods:								
RetinaNet w FPN [32]	ResNet101	$\sim 1333 \times 800$	39.1	59.1	42.3	21.8	42.7	50.2
ConRetinaNet w FPN [25]	ResNet101	$\sim 1333 \times 800$	40.1	59.6	43.5	23.4	44.2	53.3
EFGRNet [38]	ResNet101	512×512	39.0	58.8	42.3	17.8	43.6	54.5
CornerNet [26]	Hourglass104	511×511	40.5	56.5	43.1	19.4	42.7	53.9
FSAF w FPN [53]	ResNet101	$\sim 1333 \times 800$	40.9	61.5	44.0	24.0	44.2	51.3
RPDet w FPN [50]	ResNet101	$\sim 1333 \times 800$	41.0	62.9	44.3	23.6	44.1	51.7
FCOS w FPN [45]	ResNet101	$\sim 1333 \times 800$	41.5	60.7	45.0	24.4	44.8	51.6
HSD [3]	ResNet101	768×768	42.3	61.2	46.9	22.8	47.3	55.9
Two-Stage Methods:								
FPN [31]	ResNet101	$\sim 1000 \times 600$	36.2	59.1	39.0	18.2	39.0	48.2
Libra R-CNN w FPN [39]	ResNet101	$\sim 1333 \times 800$	41.1	62.1	44.7	23.4	43.7	52.5
Grid R-CNN w FPN [36]	ResNet101	$\sim 1333 \times 800$	41.5	60.9	44.5	23.3	44.9	53.1
Grid R-CNN Plus w FPN [37]	ResNet101	$\sim 1333 \times 800$	42.0	60.5	45.6	23.4	45.2	53.2
LIP w FPN [14]	ResNet101	$\sim 1333 \times 800$	42.0	64.3	45.8	24.7	45.2	52.3
Auto-FPN [49]	ResNet101	$\sim 1333 \times 800$	42.5	-	-	-	-	-
TridentNet [28]	ResNet101	$\sim 1333 \times 800$	42.7	63.6	46.5	23.9	46.6	56.6
Cascade R-CNN w FPN [2]	ResNet101	$\sim 1333 \times 800$	42.8	62.1	46.3	23.7	45.5	55.2
D2Det (ours) w FPN	ResNet101	$\sim 1333 \times 800$	**45.4**	**64.0**	**49.5**	**25.8**	**48.7**	**58.1**
DCN v2 [54]	ResNet101-deform v2	$\sim 1333 \times 800$	44.0	65.9	48.1	23.2	47.7	59.6
D2Det (ours)	ResNet101-deform v2	$\sim 1333 \times 800$	**47.4**	**65.9**	**51.7**	**27.2**	**50.4**	**61.3**
D2Det* (ours)	ResNet101-deform v2		50.1	69.4	54.9	32.7	52.7	62.1

State-of-the-art object detection comparison on COCO test-dev.

2 Experiments

Methods	Backbone	Time	Mask AP	AP@0.5	AP@0.75
MNC [9]	ResNet101	-	24.6	44.3	24.8
FCIS [30]	ResNet101	-	29.2	49.5	-
MaskLab [7]	ResNet101	-	35.4	57.4	37.4
Mask R-CNN [19]	ResNet101	116	35.7	58.0	37.8
PANet [34]	ResNet50	-	36.6	58.0	39.3
MS R-CNN [23]	ResNet101	116	38.3	58.8	41.5
Cascade Mask R-CNN [6]	ResNet101	156	38.4	60.2	41.4
HTC [6]	ResNet101	339	39.7	**61.8**	43.1
D2Det (Ours)	ResNet101	168	**40.2**	61.5	**43.7**

Table 5: State-of-the-art instance segmentation comparison (with a single model performance) in Mask AP on MS COCO `test-dev`. Other than MNC, FCIS and MaskLab, all methods employ FPN. The speed of all the methods is reported on Titan Xp. Our D2Det provides a two-fold speedup over HTC, while achieving a 40.2 mask AP.

Figure 6: Instance segmentation results of our D2Det on COCO `test-dev` (top row) and iSAID test (bottom row).

2 Experiments

Baseline	DLR	DRP	AP	AP@0.5	AP@0.75
✓			38.0	59.2	41.5
✓	✓		41.5	59.6	44.8
✓		✓	39.3	61.4	42.2
✓	✓	✓	42.7	61.5	46.3

Table 2: Impact of integrating our dense local regression (DLR) and discriminative RoI pooling (DRP) into the baseline, on MS COCO `minival`.

Method	AP	AP@0.5	AP@0.75	AP_s	AP_m	AP_l
Grid R-CNN	39.6	58.3	42.4	22.6	43.8	51.5
Grid R-CNN Plus	40.2	58.4	43.4	22.7	44.1	53.1
Our DLR alone	41.5	59.6	44.8	23.3	44.9	55.2
Ours Final: D2Det	42.7	61.5	46.3	24.5	46.2	56.9

Table 3: Comparisons of our dense local regression (DLR) with keypoint-based localization utilized in Grid R-CNN and Grid R-CNN Plus. Our DLR alone provides superior results, especially for large objects.