Beyond RetinaNet and Mask R-CNN

Gang Yu
yugang@megvii.com

Face++ 旷视

Outline

- Modern Object detectors
  - One Stage detector vs Two-stage detector
- Challenges
  - Backbone
  - Head
  - Scale
  - Batch Size
  - Crowd
- Conclusion

Face++ 旷视

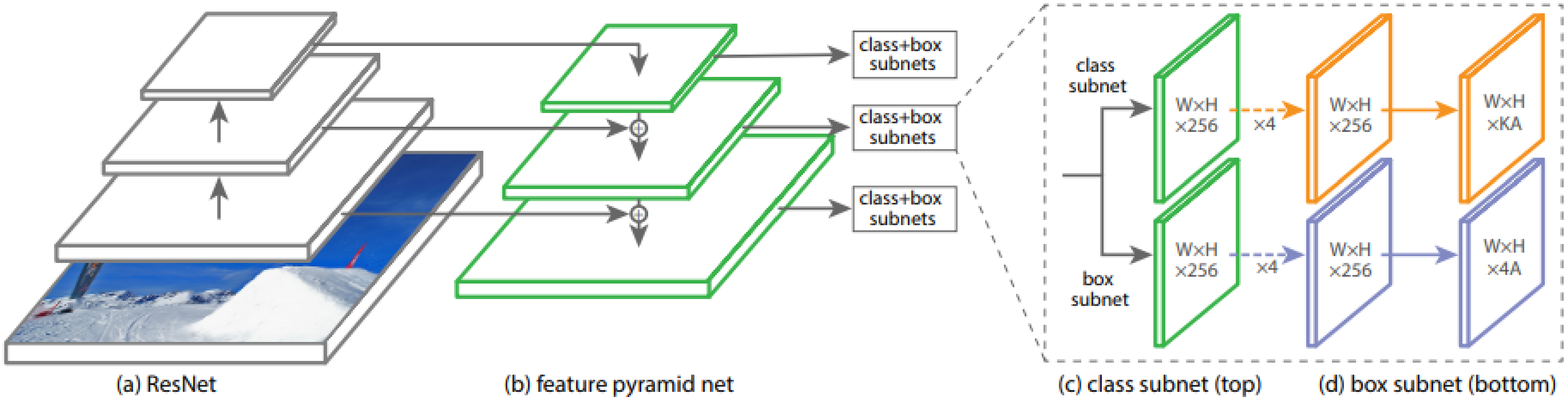
Modern Object detectors

Backbone → Head → Postprocess NMS

- Modern object detectors
  - RetinaNet
    - f1-f7 for backbone,   f3-f7 with 4 convs for head
  - FPN with ROIAlign
    - f1-f6 for backbone, two fcs for head
  - Recall vs localization
    - One stage detector: Recall is high but compromising the localization ability
    - Two stage detector: Strong localization ability

One Stage detector: RetinaNet

- FPN Structure
- Focal loss

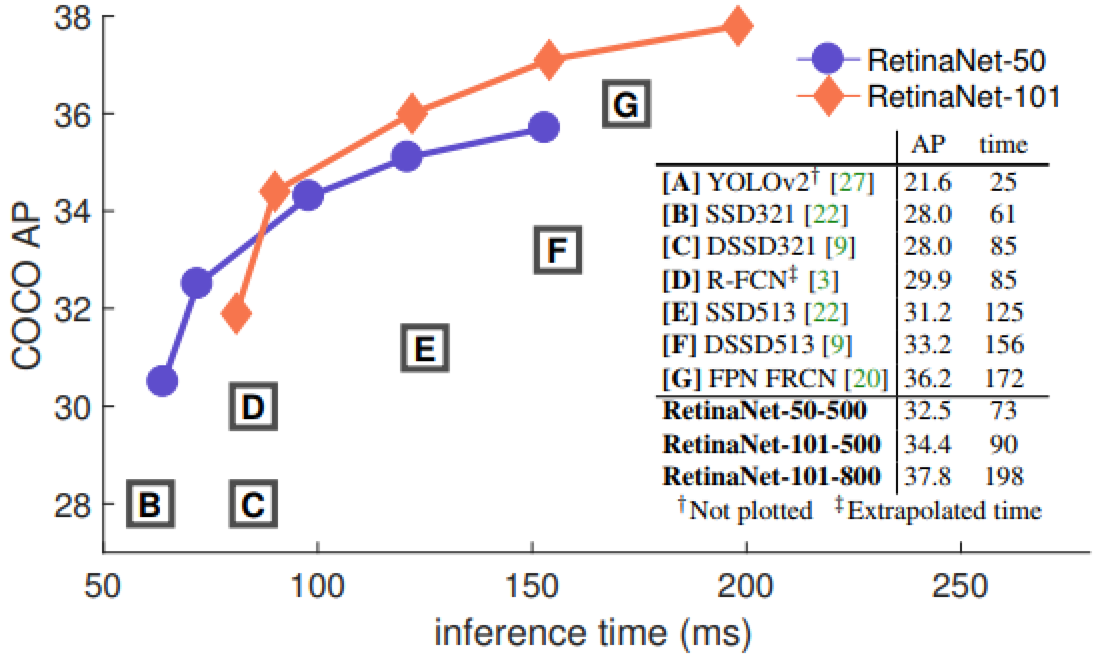


Face++ 旷视

Focal Loss for Dense Object Detection， Lin etc, ICCV 2017 Best student paper

One Stage detector: RetinaNet

- FPN Structure
- Focal loss



	AP	time
[A] YOLOv2† [27]	21.6	25
[B] SSD321 [22]	28.0	61
[C] DSSD321 [9]	28.0	85
[D] R-FCN‡ [3]	29.9	85
[E] SSD513 [22]	31.2	125
[F] DSSD513 [9]	33.2	156
[G] FPN FRCN [20]	36.2	172
RetinaNet-50-500	32.5	73
RetinaNet-101-500	34.4	90
RetinaNet-101-800	37.8	198

†Not plotted ‡Extrapolated time

Focal Loss for Dense Object Detection， Lin etc, ICCV 2017 Best student paper

Two-Stage detector: FPN/Mask R-CNN

- FPN Structure
- ROIAlign

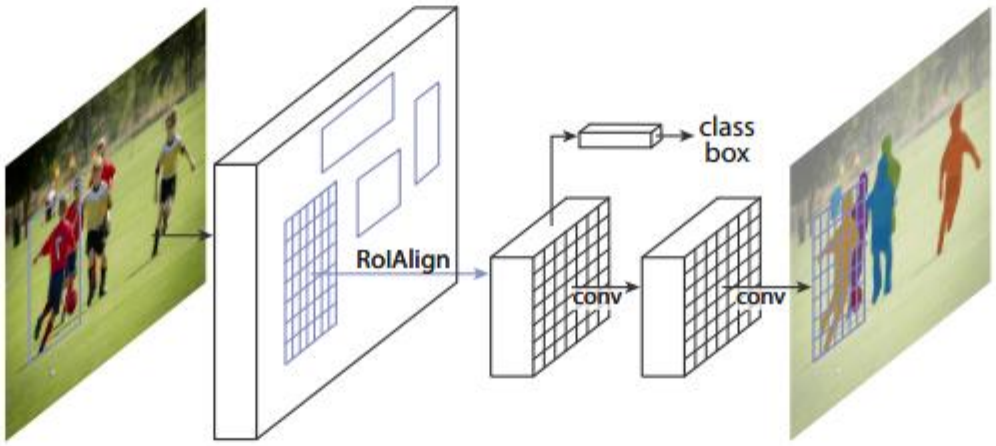


Figure 1. The **Mask R-CNN** framework for instance segmentation.

Face++ 旷视

Mask R-CNN， He etc, ICCV 2017 Best paper

What is next for object detection?

- The pipeline seems to be mature
- There still exists a large gap between existing state-of-arts and product requirements
- The devil is in the detail

Challenges Overview

- Backbone
- Head
- Scale
- Batch Size
- Crowd

Backbone → Head → Postprocess NMS

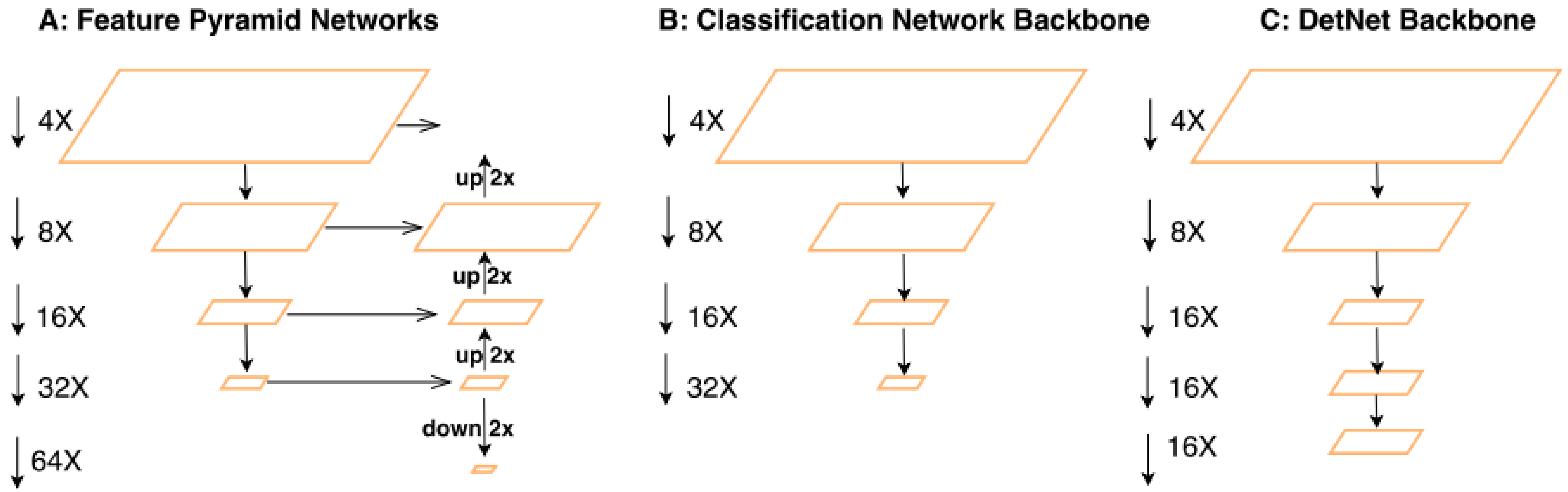
Face++ 旷视

Challenges - Backbone

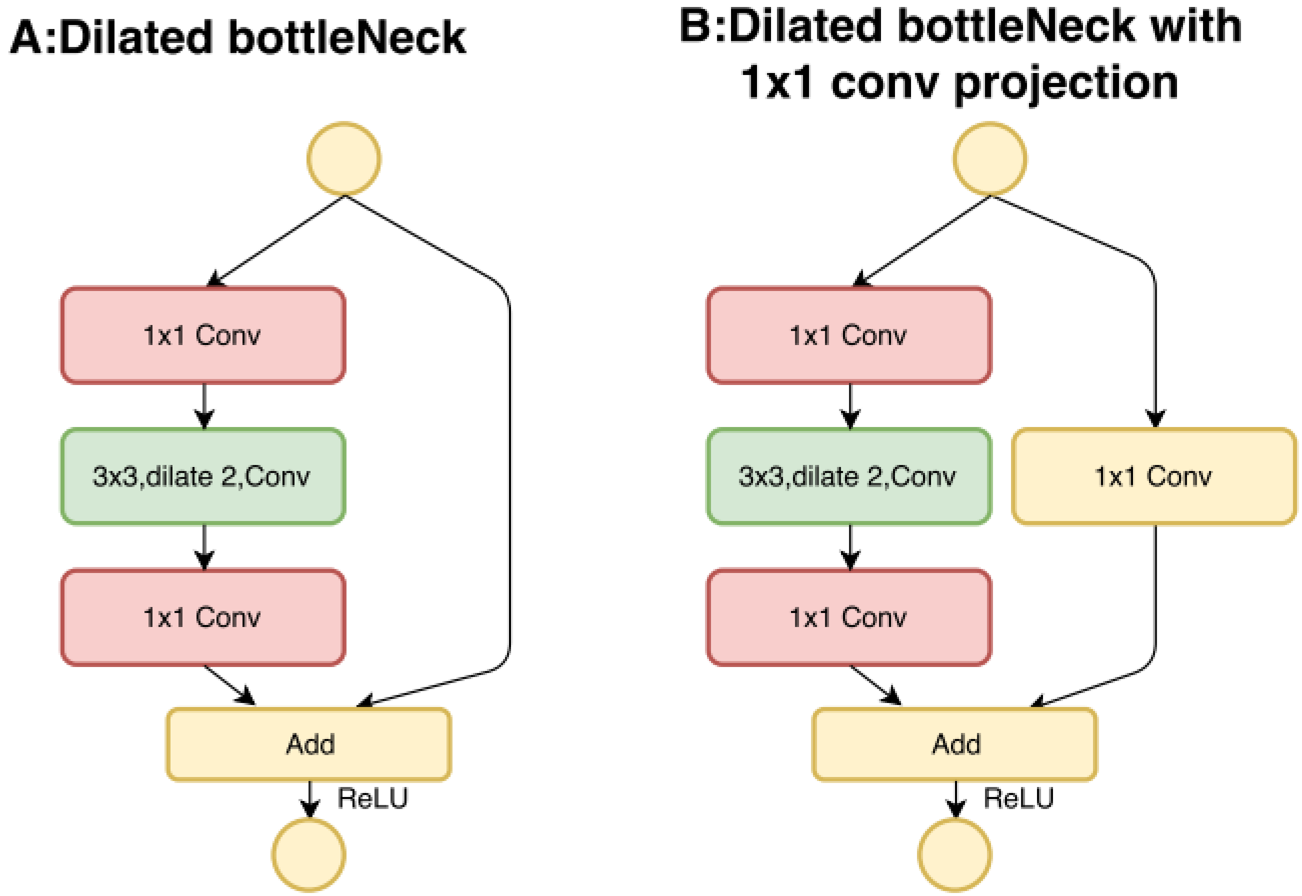
- Backbone network is designed for classification task but not for localization task
  - Receptive Field vs Spatial resolution
- Only f1-f5 is pretrained but randomly initializing f6 and f7 (if applicable)

Backbone - DetNet

- DetNet: A Backbone network for Object Detection, Li etc, 2018, https://arxiv.org/pdf/1804.06215.pdf

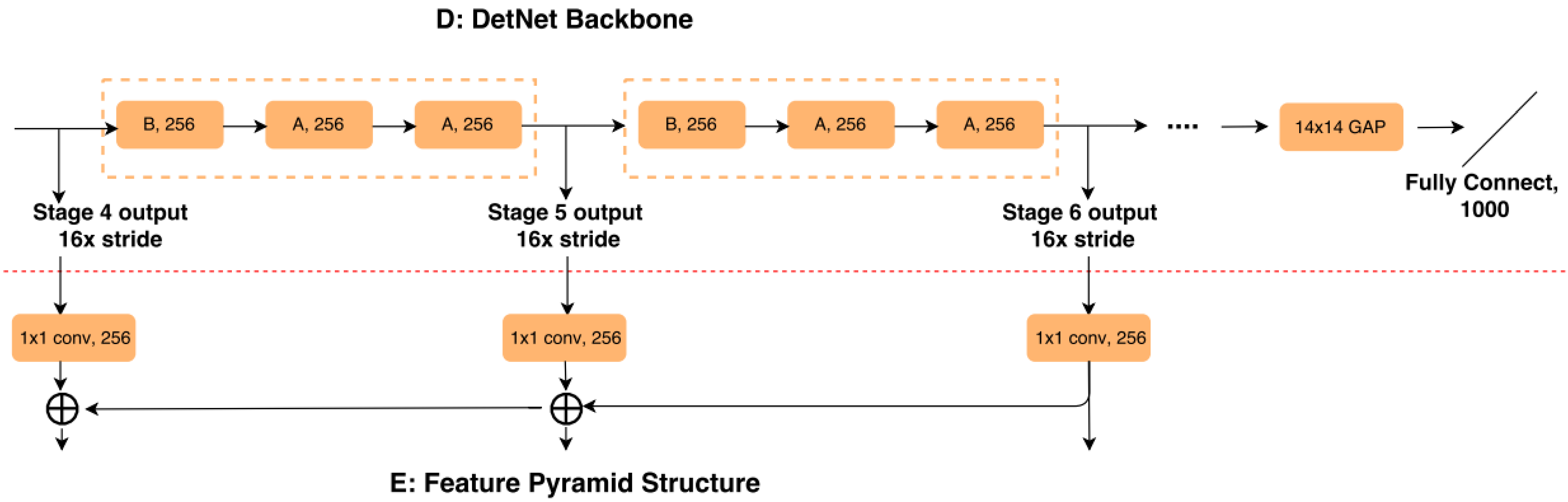


Backbone - DetNet



Face++ 旷视

Backbone - DetNet



Face++ 旷视

Backbone - DetNet

bacbone	Classification		FPN on COCO minival						FPN on COCO test-dev					
	Err	FLOPs	mAP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	mAP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
D-59	23.5	4.8G	**40.2**	61.7	43.7	23.9	43.2	52.0	**40.3**	62.1	43.8	23.6	42.6	50.0
R-62	23.4	4.7G	38.8	60.6	42.4	22.6	41.6	51.6	39.0	61.0	42.3	21.9	41.2	49.7
R-50	24.1	3.8G	37.9	60.0	41.2	22.9	40.6	49.2	38.4	60.4	41.6	22.5	40.7	47.9
D-101	23.0	7.9G	**41.9**	62.8	45.7	25.4	45.2	55.1	**42.2**	63.2	45.8	24.5	44.8	53.1
R-101	22.9	7.8G	39.9	62.0	43.7	24.1	43.4	52.0	40.3	62.5	44.0	23.3	43.1	50.6

Table 1. Comparison of 'D' DetNet and 'R' ResNet. We report both results on ImageNet classification (Top1 Error) and FPN COCO detection. Results validate that DetNet is more suitable for object detection. Keeping same model size, DetNet consistently outperform ResNet.

Backbone - DetNet

Models	scales	mAP	AP_{50}	AP_{60}	AP_{70}	AP_{80}	AP_{85}
ResNet-50	over all scales	37.9	60.0	55.1	47.2	33.1	22.1
	small	22.9	40.1	35.5	28.0	17.5	10.4
	middle	40.6	63.9	59.0	51.2	35.7	23.3
	large	49.2	72.2	68.2	60.8	46.6	34.5
DetNet-59	over all scales	40.2	61.7	57.0	49.6	36.2	25. 8
	small	23.9	41.8	36.8	29.8	17.7	10.5
	middle	43.2	65.8	61.2	53.6	39.9	27.3
	large	52.0	73.1	69.5	63	51.4	40.0

Models	scales	mAR	AR_{50}	AR_{60}	AR_{70}	AR_{80}	AR_{85}
ResNet-50	over all scales	52.8	80.5	74.7	64.3	46.8	34.2
	small	35.5	60.0	53.8	43.3	28.7	18.7
	middle	56.0	84.9	79.2	68.7	50.5	36.2
	large	67.0	95.0	90.9	80.3	63.1	50.2
DetNet-59	over all scales	56.1	83.1	77.8	67.6	51.0	38.9
	small	39.2	66.4	59.4	47.3	29.5	19.6
	middle	59.5	87.4	82.5	72.6	55.6	41.2
	large	70.1	95.4	91.8	82.9	69.1	56.3

Backbone - DetNet

Models	Backbone	mAP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
SSD513 [3]	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513 [3,37]	ResNet-101	33.2	53.3	35.2	13.0	35.4	51.1
Faster R-CNN +++ [11]	ResNet-101	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN G-RMI [2] [38]	Inception-ResNet-v2	34.7	55.5	36.7	13.5	38.1	52.0
RetinaNet [4]	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2
FPN [33]	ResNet-101	37.3	59.6	40.3	19.8	40.2	48.8
FPN	**DetNet-59**	**40.3**	62.1	43.8	23.6	42.6	50.0

Models	Backbone	mAP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
MNC [39]	ResNet-101	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [40] + OHEM [41]	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
FCIS+++ [40] +OHEM	ResNet-101-C5-dilated	33.6	54.5	-	-	-	-
Mask R-CNN [33]	ResNet-101	35.7	58.0	37.8	15.5	38.1	52.4
Mask R-CNN	**DetNet-59**	**37.1**	60.0	39.6	18.6	39.0	51.3

Face++ 旷视

Challenges - Head

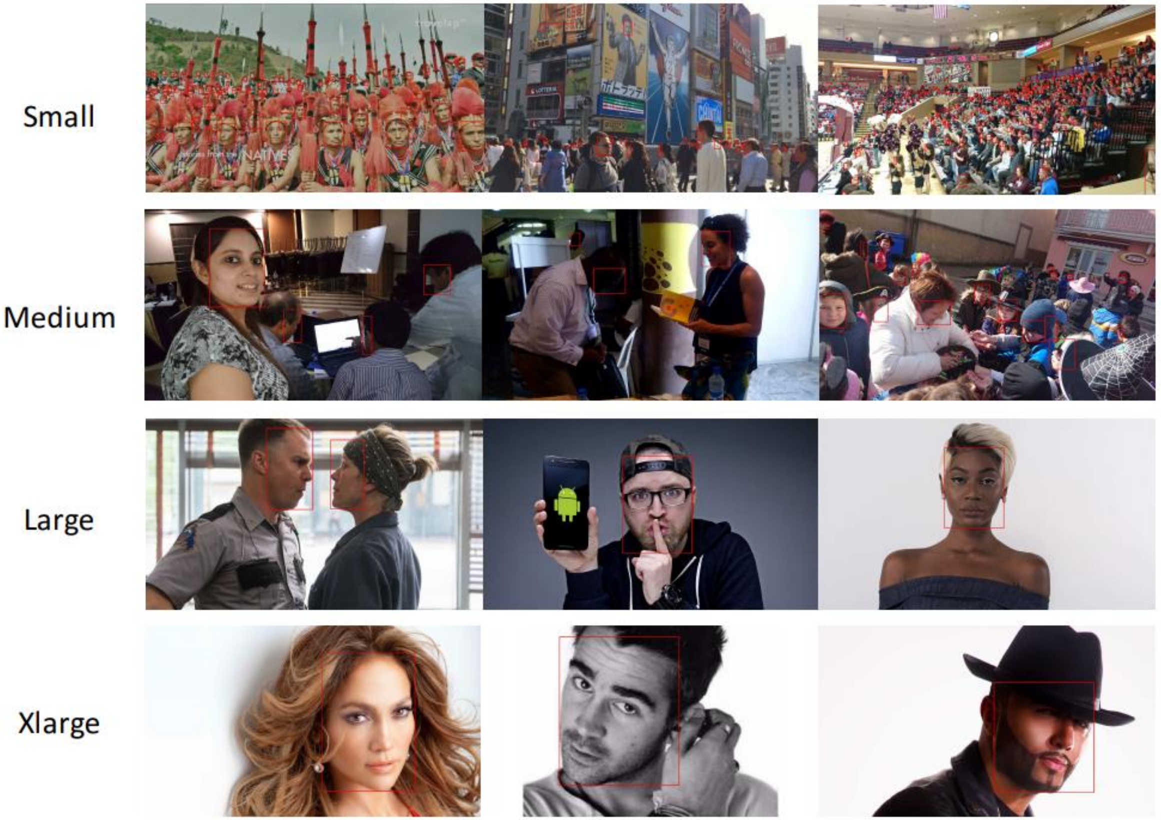
- Speed is significantly improved for the two-stage detector
  - RCNN - > Fast RCNN -> Faster RCNN - > RFCN
- How to obtain efficient speed as one stage detector like YOLO, SSD?
  - Small Backbone
  - Light Head

Head – Light head RCNN

- Light-Head R-CNN: In Defense of Two-Stage Object Detector, 2017, https://arxiv.org/pdf/1711.07264.pdf

Challenges - Scale

- Scale variations is extremely large for object detection



Face++ 旷视

Challenges - Scale

- Scale variations is extremely large for object detection
- Previous works
  - Divide and Conquer: SSD, DSSD, RON, FPN, …
    - Limited Scale variation
  - Scale Normalization for Image Pyramids, Singh etc, CVPR2018
    - Slow inference speed
- How to address extremely large scale variation without compromising inference speed?

Scale - SFace

- SFace: An Efficient Network for Face Detection in Large Scale Variations, 2018, http://cn.arxiv.org/pdf/1804.06559.pdf

Challenges - Batchsize

- Small mini-batchsize for general object detection
  - 2 for R-CNN, Faster RCNN
  - 16 for RetinaNet, Mask RCNN
- Problem with small mini-batchsize
  - Long training time
  - Insufficient BN statistics
  - Inbalanced pos/neg ratio

Batchsize – MegDet

- MegDet: A Large Mini-Batch Object Detector, CVPR2018, https://arxiv.org/pdf/1711.07240.pdf

Challenges - Crowd

- NMS is a post-processing step to eliminate multiple responses on one object instance
  - Reasonable for mild crowdness like COCO and VOC
  - Will Fail in the case when the objects are in a crowd

Figure 1. Illustrative examples from different human dataset benchmarks. The images inside the green, yellow, blue boxes are from the COCO [17], Caltech [6], and CityPersons [31] datasets, respectively. The images from the second row inside the red box are from our CrowdHuman benchmark with full body, visible body, and head bounding box annotations for each person.

Crowd - CrowdHuman

- CrowdHuman: A Benchmark for Detecting Human in a Crowd, 2018, https://arxiv.org/pdf/1805.00123.pdf

Introduction to Face++ Detection Team

- Category-level Recognition
  - Detection
    - Face Detection:
      - FAN: https://arxiv.org/pdf/1711.07246.pdf
      - Sface: https://arxiv.org/pdf/1804.06559.pdf
    - Human Detection:
      - Repulsion loss: https://arxiv.org/abs/1711.07752
      - CrowdHuman: https://arxiv.org/pdf/1805.00123.pdf
    - General Object Detection:
      - Light Head: https://arxiv.org/pdf/1711.07264.pdf https://github.com/zengarden/light_head_rcnn
      - MegDet: https://arxiv.org/pdf/1711.07240.pdf
      - DetNet: https://arxiv.org/pdf/1804.06215.pdf
  - Segmentation
    - Large Kernel Matters: https://arxiv.org/pdf/1703.02719.pdf
    - DFN: https://arxiv.org/pdf/1804.09337.pdf
  - Skeleton:
    - CPN: https://arxiv.org/pdf/1711.07319.pdf
    - https://github.com/chenyilun95/tf-cpn

Face++ 旷视

Thanks

Face++ 旷视