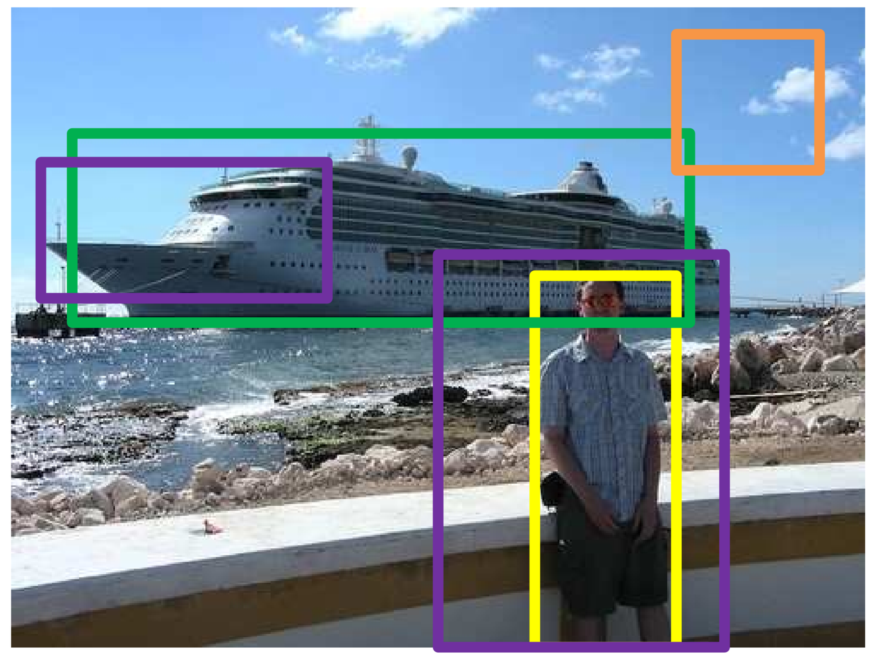
Cascade R-CNN: Delving into High Quality Object Detection

Zhaowei Cai   Nuno Vasconcelos

University of California San Diego, Statistical Visual Computing Laboratory

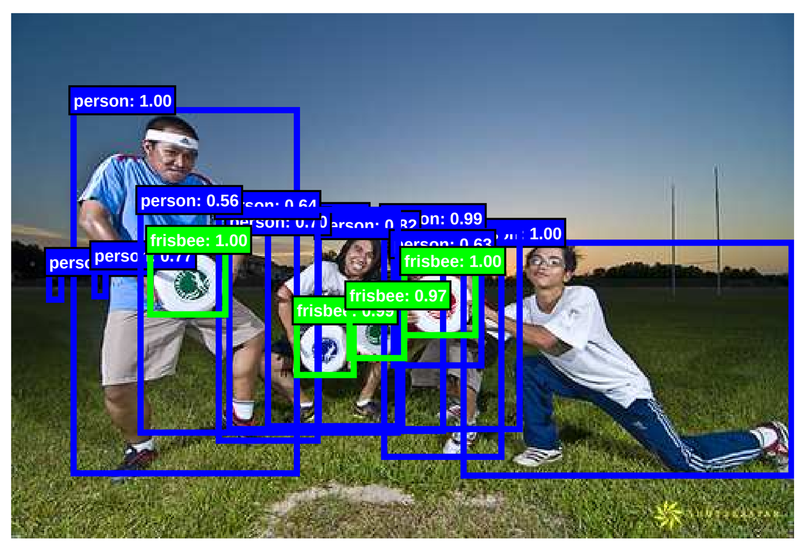
I. Object Detection

- Positive/Negative Definition
  - It is ambiguous.

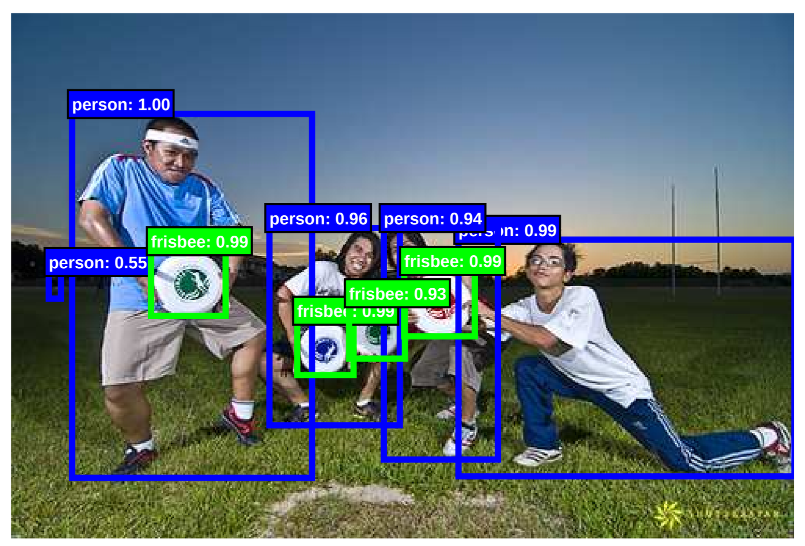


positive??
positive??
positive??
negative??

- An intersection over union (IoU) threshold is used as empirical solution (typically 0.5).
- This usually produces noisy/low-quality detection.



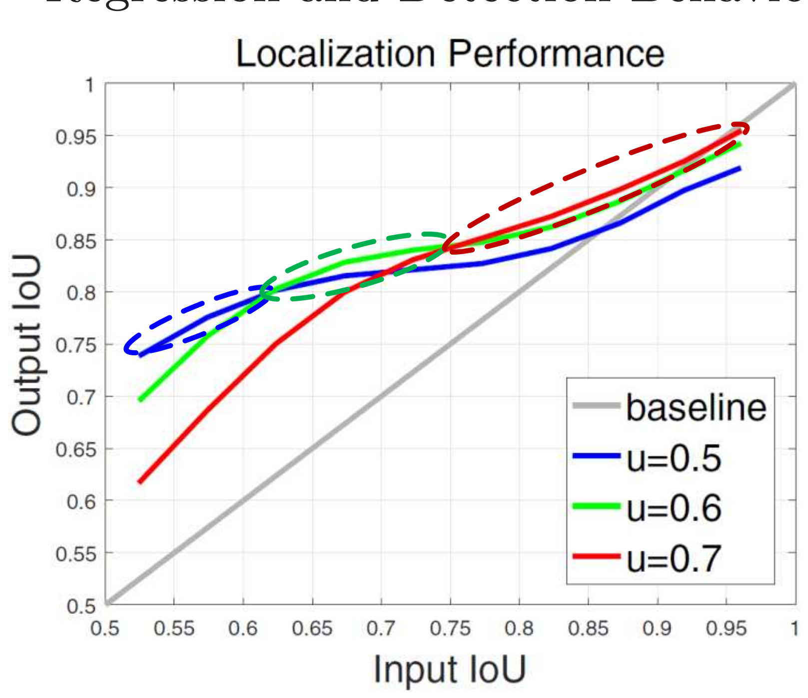
Low Quality Detection

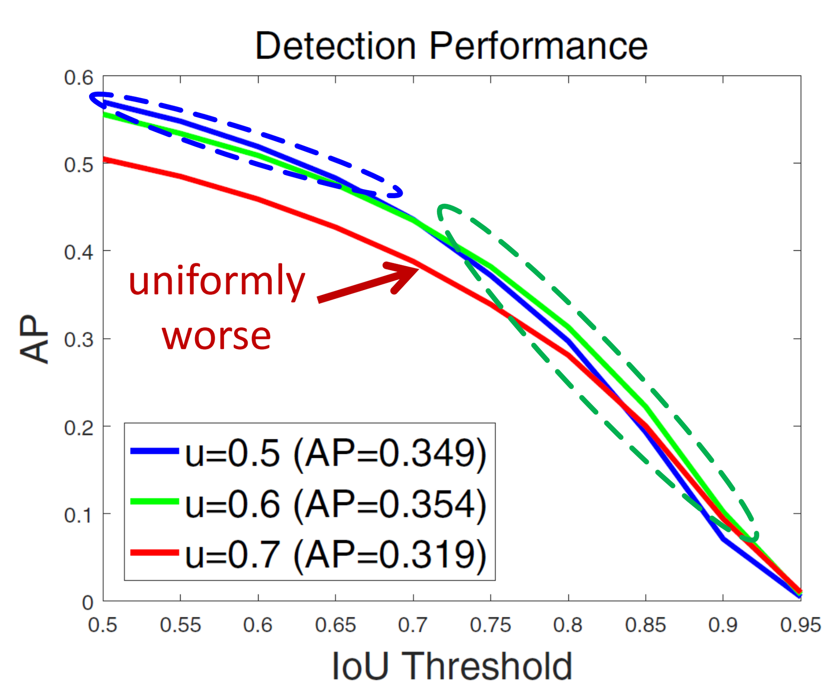


High Quality Detection

II. High Quality Object Detection

- Regression and Detection Behavior

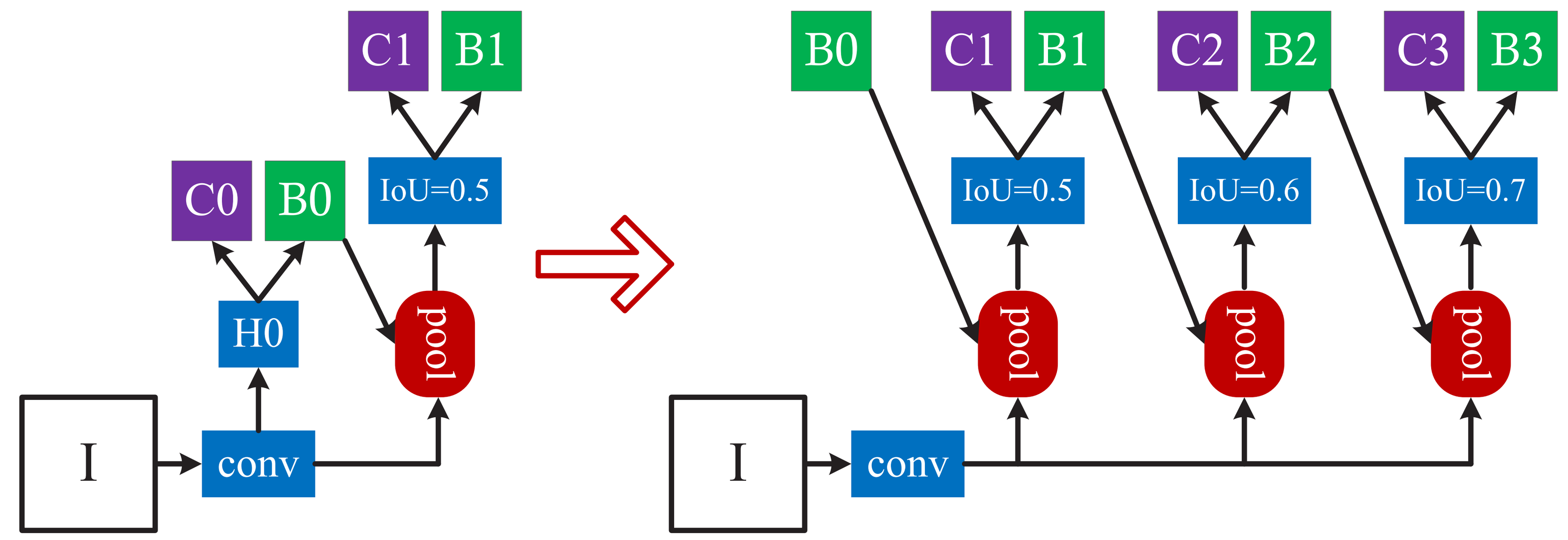




- To produce a high quality detector, it does not suffice to simply increase IoU threshold during training.
- Two main factors are responsible for this:
  - Overfitting during training, due to the exponentially vanishing positive samples.
  - Inference-time quality mismatch between the detector and its input hypotheses.

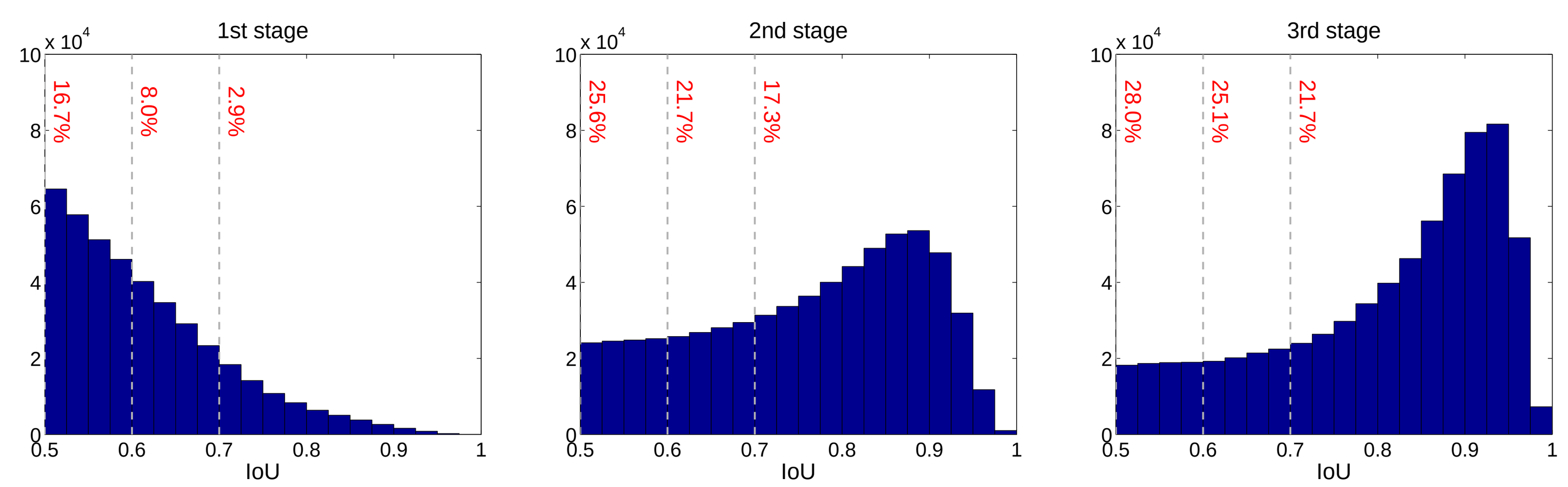
III. Cascade R-CNN

- Multi-stage Detection Framework
  - It is a multi-stage extension of the R-CNN, where detector stages deeper into the cascade are sequentially more selective against close false positives.

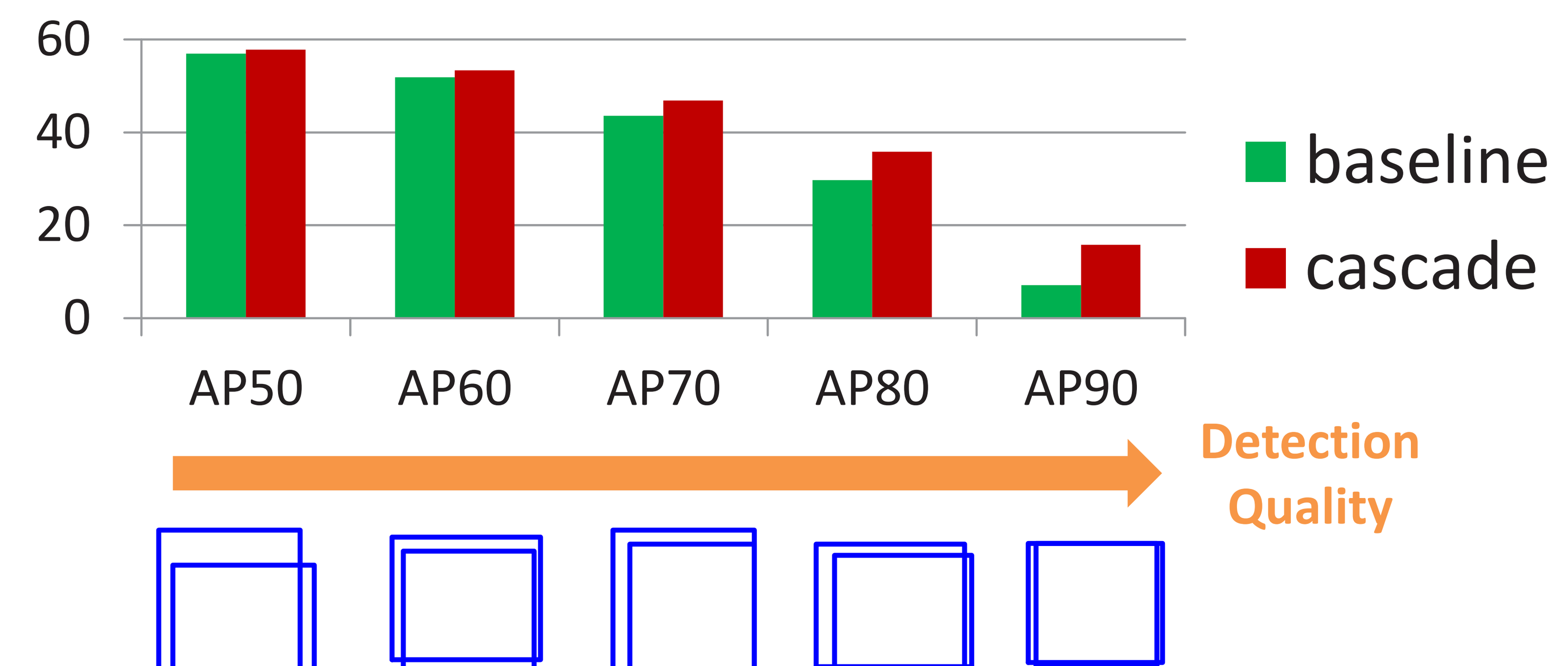


Faster R-CNN   Cascade R-CNN

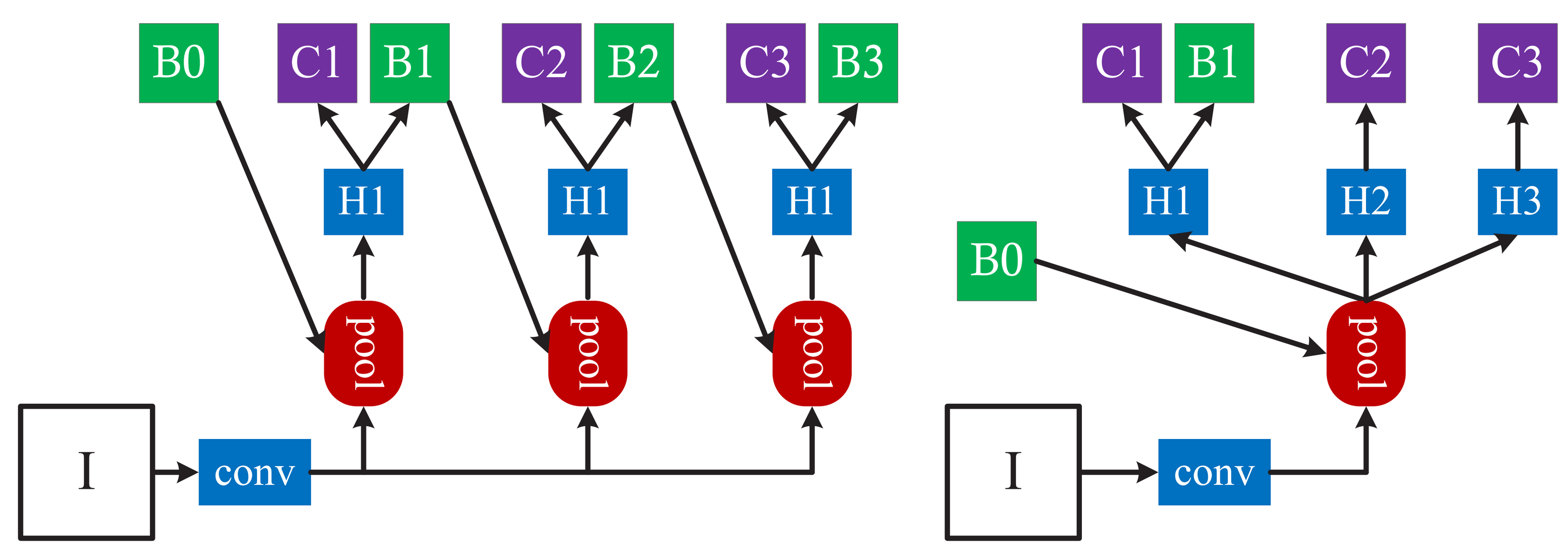
- Why Cascade R-CNN?
  - reduce training overfitting.
  - reduce inference-time quality mismatch.



- Large gains for high quality detection



- Difference with Related Works



Iterative BBox at inference   Integral Loss

IV. Experimental Results

- Comparison with related works

	AP	AP_{50}	AP_{60}	AP_{70}	AP_{80}	AP_{90}
FPN+ baseline	34.9	57.0	51.9	43.6	29.7	7.1
Iterative BBox	35.4	57.2	52.1	44.2	30.4	8.1
Integral Loss	35.4	57.3	52.5	44.4	29.9	6.9
Cascade R-CNN	**38.9**	**57.8**	**53.4**	**46.9**	**35.8**	**15.8**

- Comparison with the state-of-the-art on COCO

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2
FPN	ResNet-101	36.2	59.1	39.0	18.2	39.0	48.2
G-RMI	In-ResNet-v2	34.7	55.5	36.7	13.5	38.1	52.0
Deformable R-FCN	Algn-In-ResNet	37.5	58.0	40.8	19.4	40.1	52.5
Mask R-CNN	ResNet-101	38.2	60.3	41.7	20.1	41.1	50.2
Cascade R-CNN	ResNet-101	**42.8**	**62.1**	**46.3**	**23.7**	**45.5**	**55.2**

- Generalization on multiple detectors and backbone networks

	backbone	cascade	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster R-CNN	VGG	✗	23.5	43.9	22.6	8.1	25.1	34.7
		✓	26.9	44.3	27.8	8.3	28.2	41.1
R-FCN	ResNet-50	✗	27.1	49.0	26.9	10.4	29.7	39.2
		✓	30.9	49.9	32.6	10.5	33.1	46.9
R-FCN	ResNet-101	✗	30.5	52.9	31.2	12.0	33.9	43.8
		✓	33.3	52.6	35.2	12.1	36.2	49.3
FPN+	ResNet-50	✗	36.5	59.0	39.2	20.3	38.8	46.4
		✓	40.6	59.9	44.0	22.6	42.7	52.1
FPN+	ResNet-101	✗	38.8	61.1	41.9	21.3	41.8	49.8
		✓	42.8	62.1	46.3	23.7	45.5	55.2

- Generalization on VOC

	Faster R-CNN				R-FCN			
backbone	AlexNet		VGG		RetNet-50		RetNet-101	
cascade	✗	✓	✗	✓	✗	✓	✗	✓
AP	29.4	38.9	42.9	51.2	44.8	51.8	49.4	54.2
AP_{50}	63.2	66.5	76.4	79.1	77.5	78.5	79.8	79.6
AP_{75}	23.7	40.5	44.1	56.3	46.8	57.1	53.2	59.2

- Reproducible research
  - https://github.com/zhaoweicai/cascade-rcnn

V. Conclusions

- Cascade R-CNN
  - is an effective high quality object detector;
  - is well motivated from experimental observations;
  - achieves the state-of-the-art single-model results on COCO, and can be well generalized on other datasets, e.g. VOC;
  - can be built with any two-stage object detector based on the R-CNN framework;
  - enables consistent gains on multiple baseline detectors with multiple backbone networks, and the gain is independent of the baseline strength;
  - is quite simple to implement and reproducible on multiple codebase.

CVPR 2018 SALT LAKE CITY • JUNE 18-22

Computer Vision and Pattern Recognition, 2018