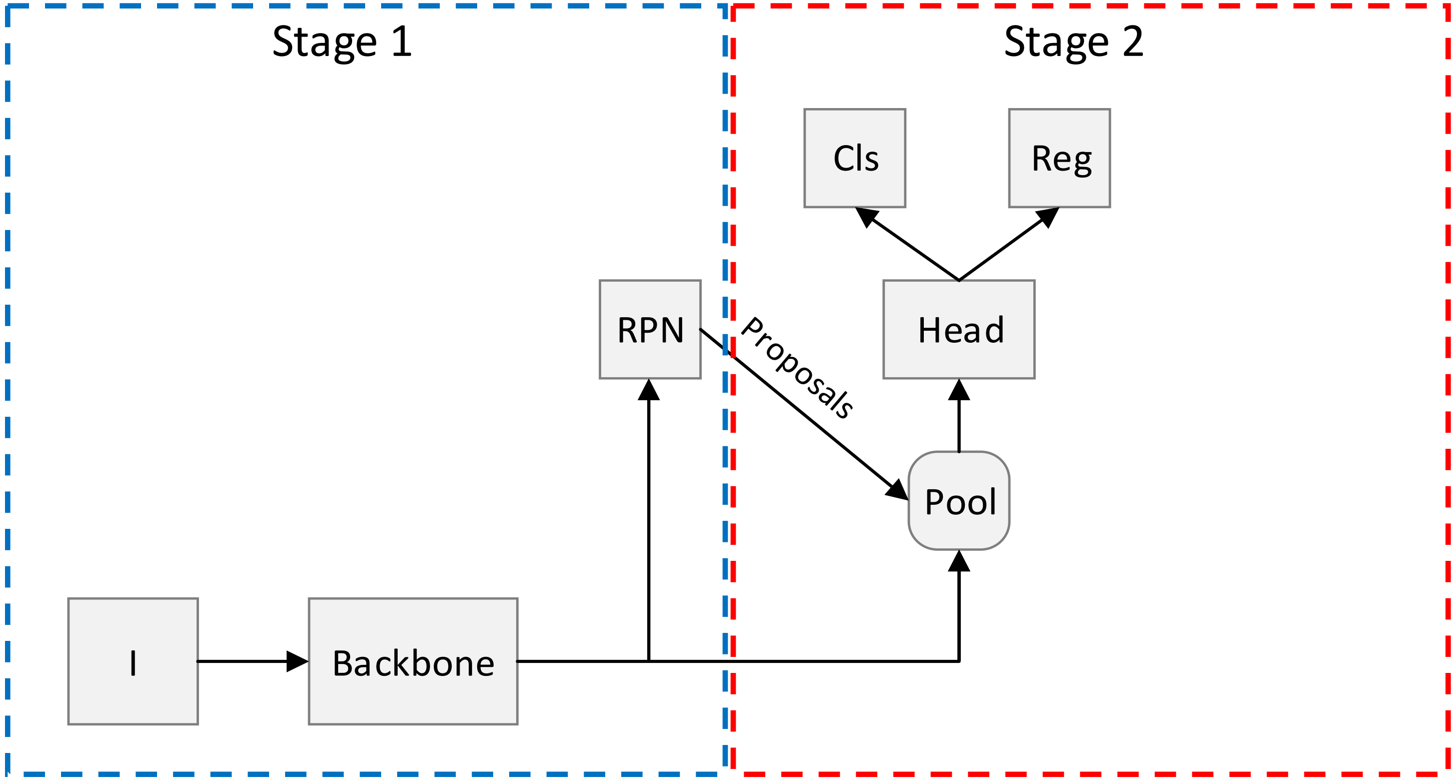
KAIST

Cascade RPN: Delving into High-Quality Region Proposal Network with Adaptive Convolution

Thang Vu, Hyunjun Jang, Pham X. Trung, Chang D. Yoo

Korea Advanced Institute of Science and Technology

Background



Faster R-CNN [1]

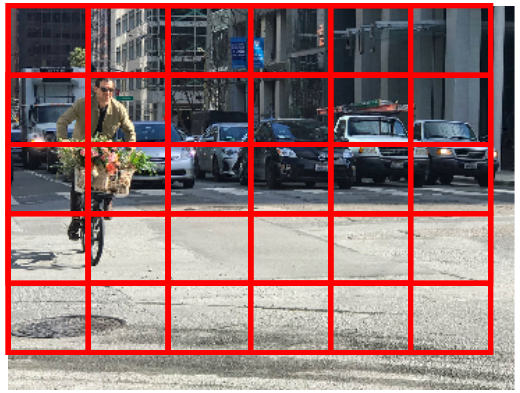
The proposed method aims to improve the RPN in stage 1

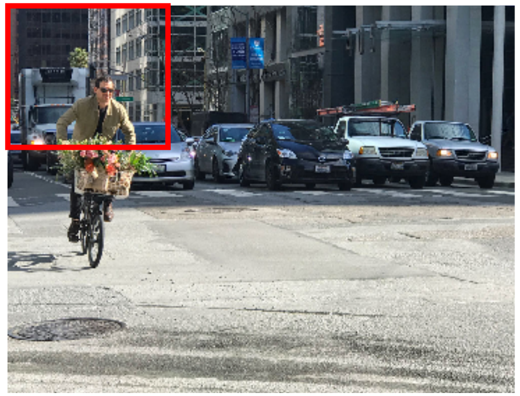
[1] Ren et al., NeuIPS 2015

Background

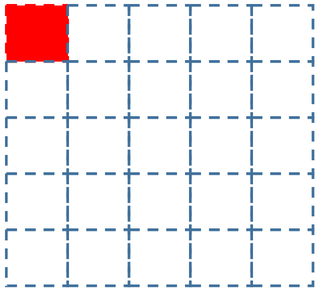
- Anchor boxes:
  - The reference for regression of RPN
  - Predefined
  - Uniformly initialized over the image
- Alignment in RPN design
  - A feature map pixel should well-align to it reference anchor boxes

  (e.g., top-left pixel should predict for top-left anchor box)



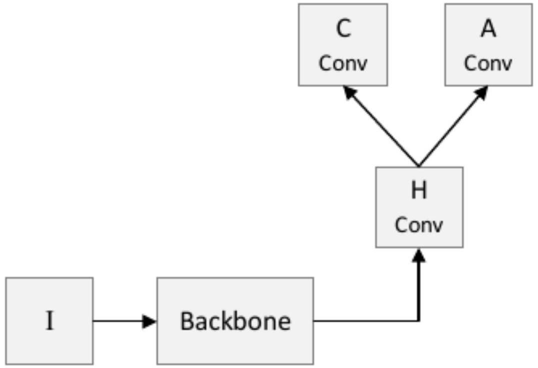


Image

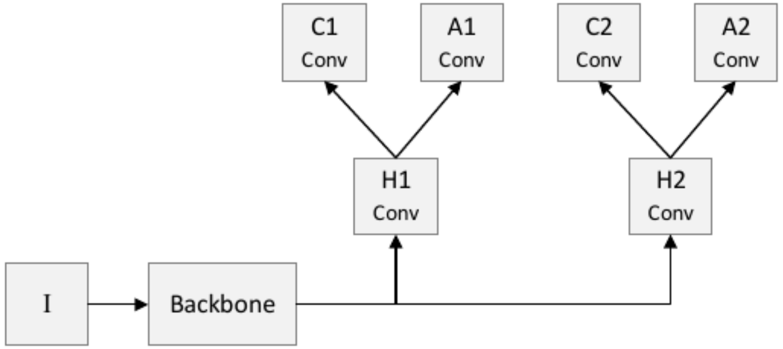


Feature

Cascade RPN



RPN [1]



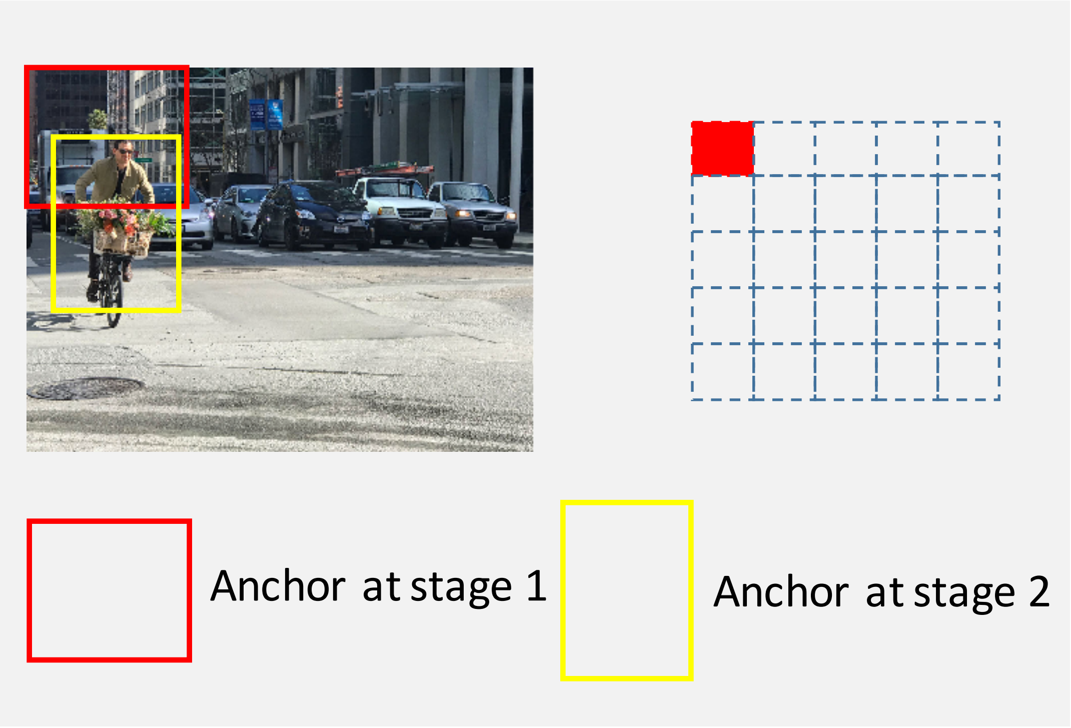
Iterative RPN [2]

- In standard RPN: Anchor is initialized uniformly using sliding window

➡ Standard conv layers can be used.

- In Iterative RPN: Anchor position and shape (after the first stage) change arbitrarily

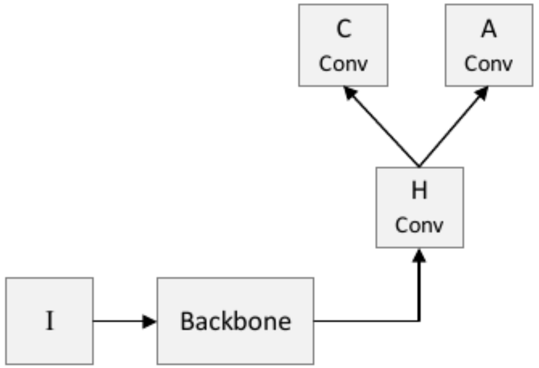
➡ Standard conv layers will break alignment between feature and anchor



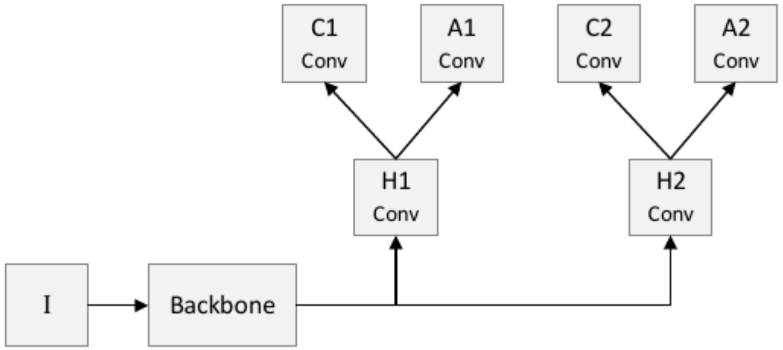
[1] Ren et al., Toward real-time object detection with RPN, NeurIPS 2015

[2] Zhong et al., Cascade region proposal and global context for deep object detection, arXiv 2018

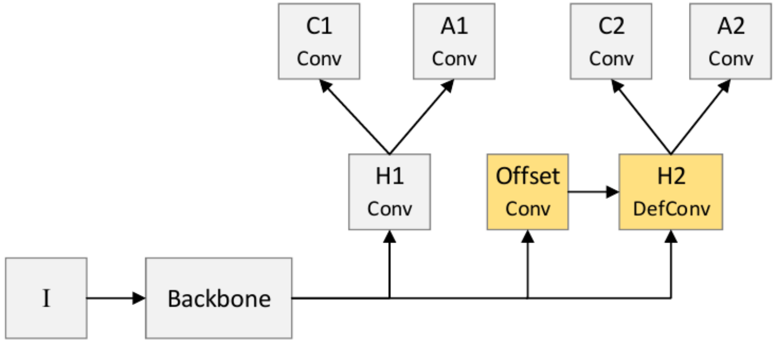
Cascade RPN



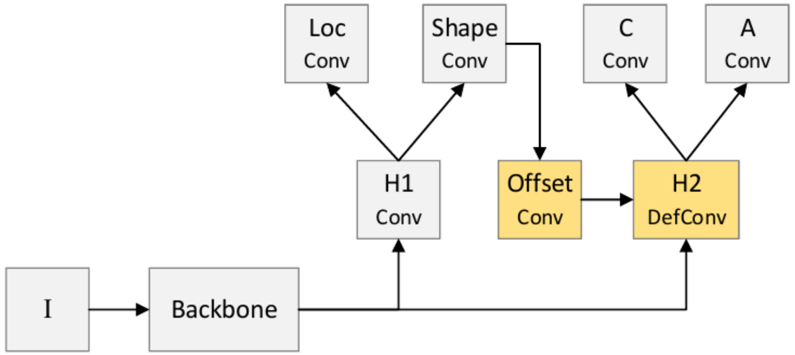
RPN [1]



Iterative RPN [2]



Iterative RPN+ [3]



GA-RPN [4]

- Deformable conv learn arbitrary feature transformation
- There is no constraint to make deformable conv produce alignment between anchor and feature

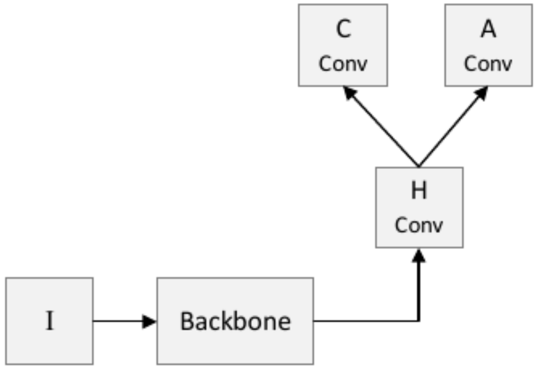
[1] Ren et al., Toward real-time object detection with RPN, NeurIPS 2015

[2] Zhong et al., Cascade region proposal and global context for deep object detection, arXiv 2018

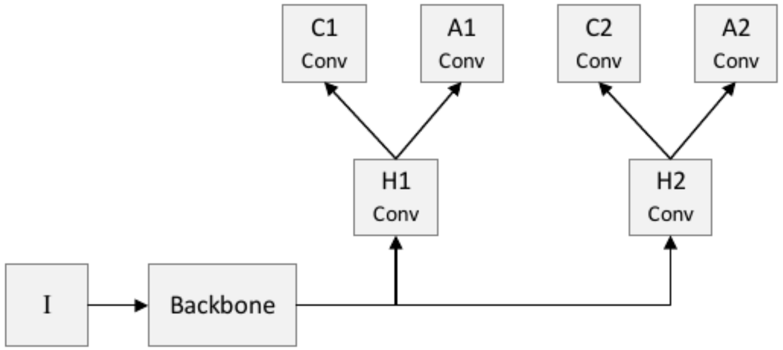
[3] Fan et al., Siamese cascaded region proposal networks for real-time visual tracking

[4] Wang et al., Region proposal by guided anchoring, CVPR 2019

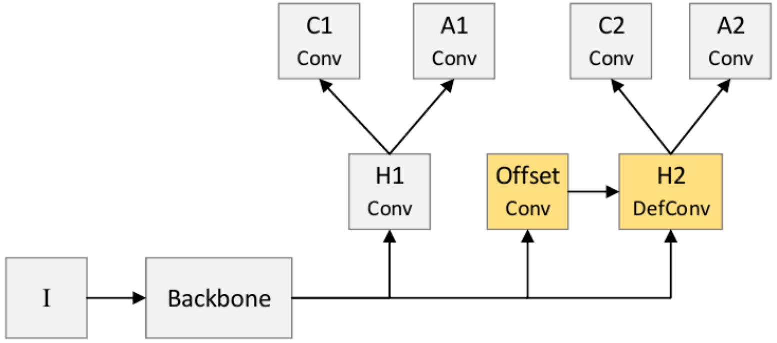
Cascade RPN



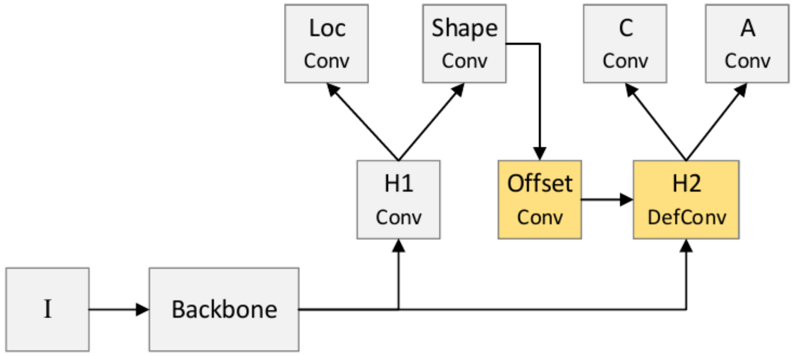
RPN [1]



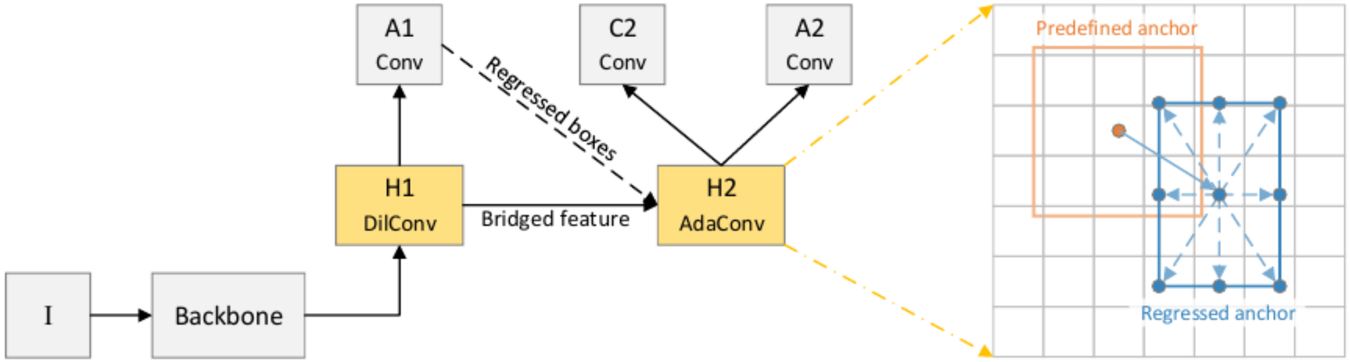
Iterative RPN [2]



Iterative RPN+ [3]



GA-RPN [4]



Cascade RPN (ours)

[1] Ren et al., Toward real-time object detection with RPN, NeurIPS 2015

[2] Zhong et al., Cascade region proposal and global context for deep object detection, arXiv 2018

[3] Fan et al., Siamese cascaded region proposal networks for real-time visual tracking

[4] Wang et al., Region proposal by guided anchoring, CVPR 2019

Adaptive Convolution

- Standard Convolution
  - Sample at regular grid $\mathbb{R}$

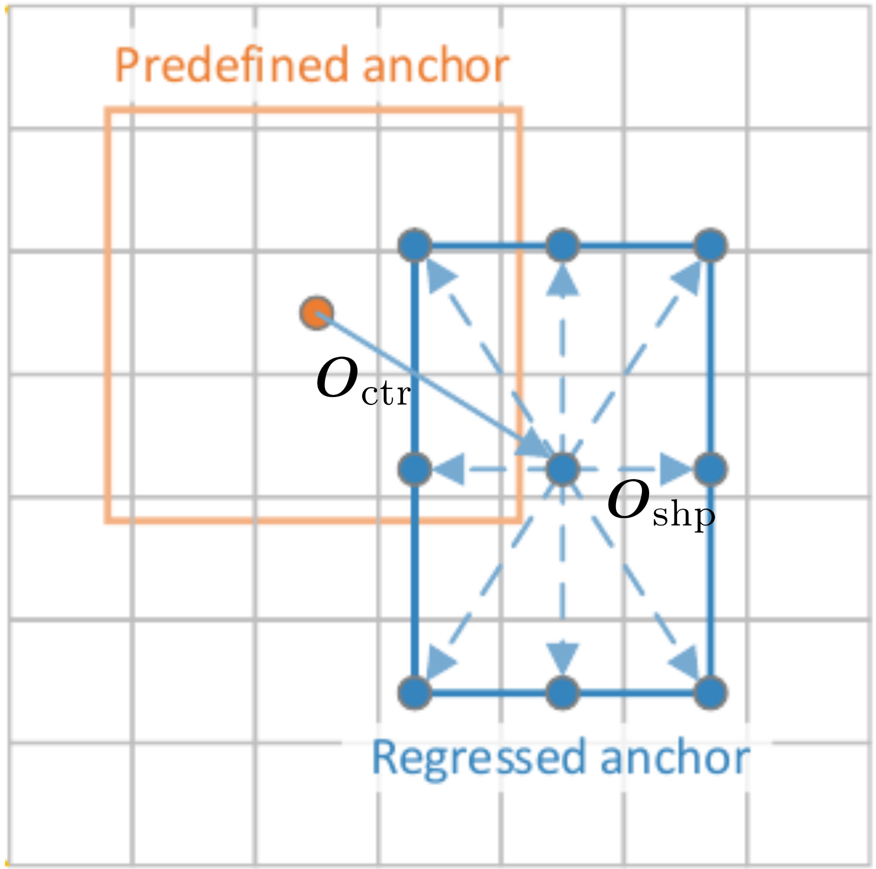
$$\mathbb{R} = \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\}$$

$$\boldsymbol{y}[\boldsymbol{p}] = \boxed{\sum_{\boldsymbol{r}\in\mathbb{R}}} \boldsymbol{w}[\boldsymbol{p}] \cdot \boldsymbol{x}[\boldsymbol{p}+\boldsymbol{r}]$$

- Adaptive Convolution
  - Sample at offset grid $\mathbb{O}$, guided by anchor

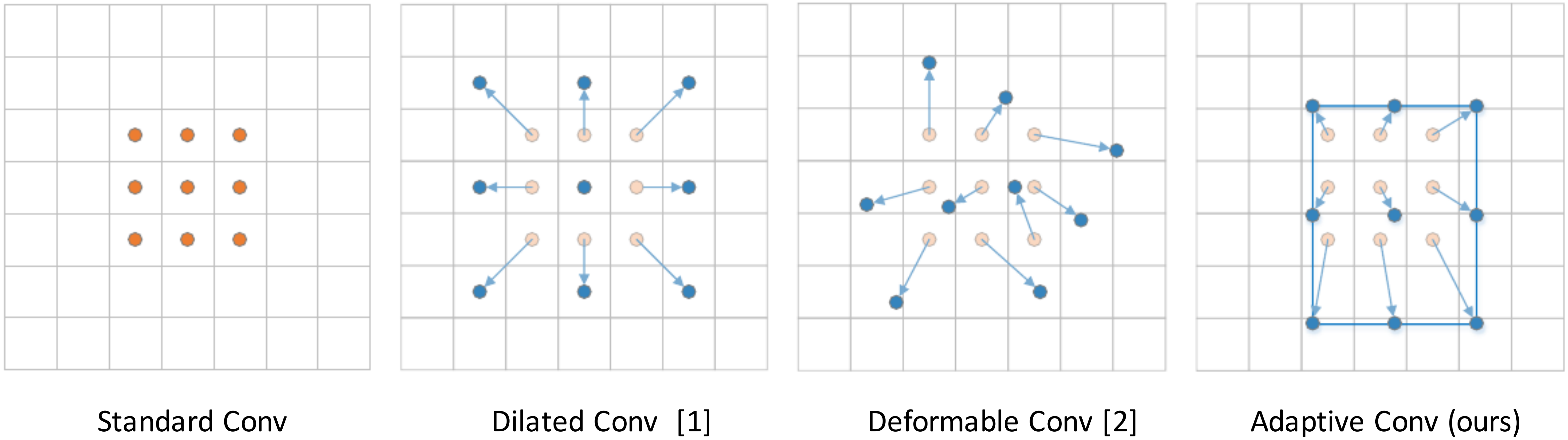
$$\boldsymbol{y}[\boldsymbol{p}] = \boxed{\sum_{\boldsymbol{o}\in\mathbb{O}}} \boldsymbol{w}[\boldsymbol{p}] \cdot \boldsymbol{x}[\boldsymbol{p}+\boldsymbol{o}]$$

$$\boldsymbol{o} = \boldsymbol{o}_{\mathrm{ctr}} + \boldsymbol{o}_{\mathrm{shp}}$$



Adaptive conv systematically maintain alignment between features and anchors!

Relation to other Convolutions



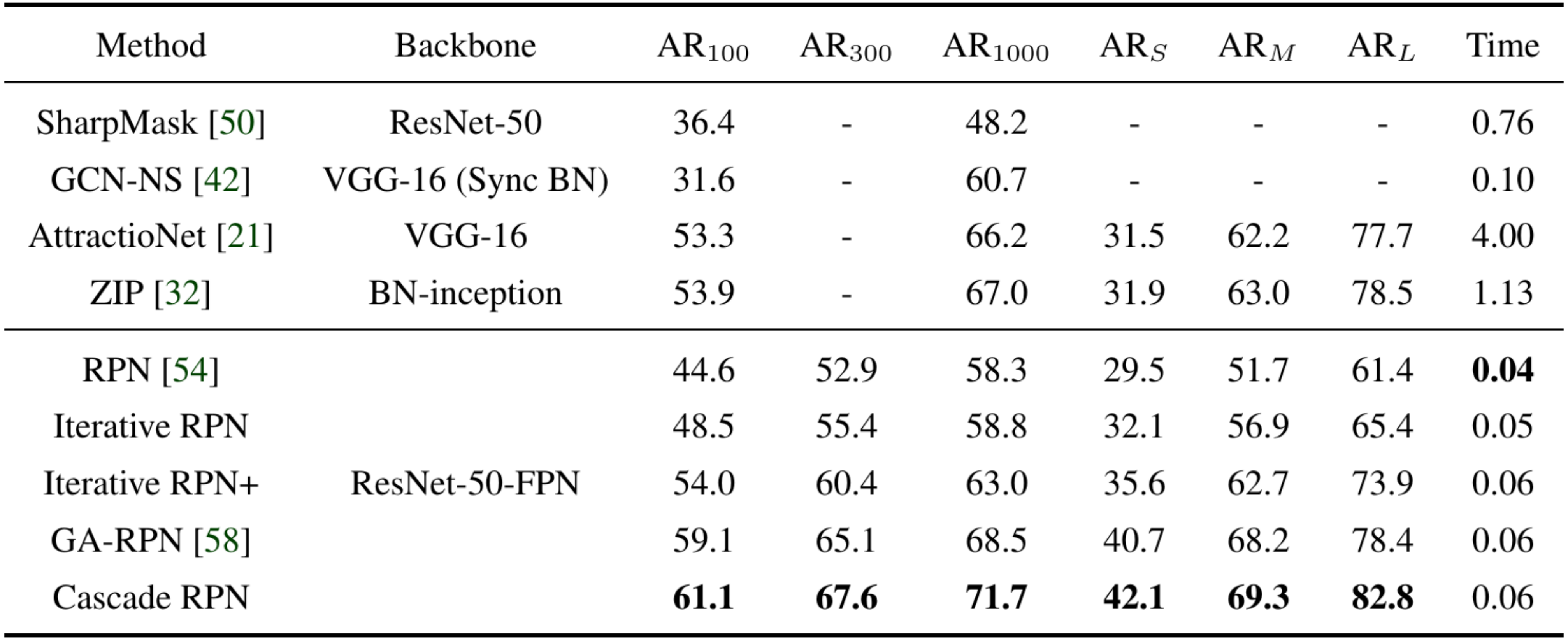
- Adaptive Conv is closely related to the others
  - Adaptive Conv becomes Dilated Conv if center offsets are 0
  - Deformable Conv becomes Adaptive Conv if offsets are deterministically derived from anchors.

[1] Yu et al., arXiv 2015
[2] Dai et al., ICCV 2017

Experiments

- Dataset: COCO2017
  - Train: 115k images
  - Val: 5k images
  - Test-dev: 20k images
- Default model:
  - Backbone: ResNet50-FPN
  - Without bells and whistles
  - Train 14 hours on 8 V100 GPUs
- Evaluation metric:
  - Average Recall (AR) for Region Proposal performance
  - Average Precision (AP) for Detection performance
  - Runtime is measured on a single V100

Results



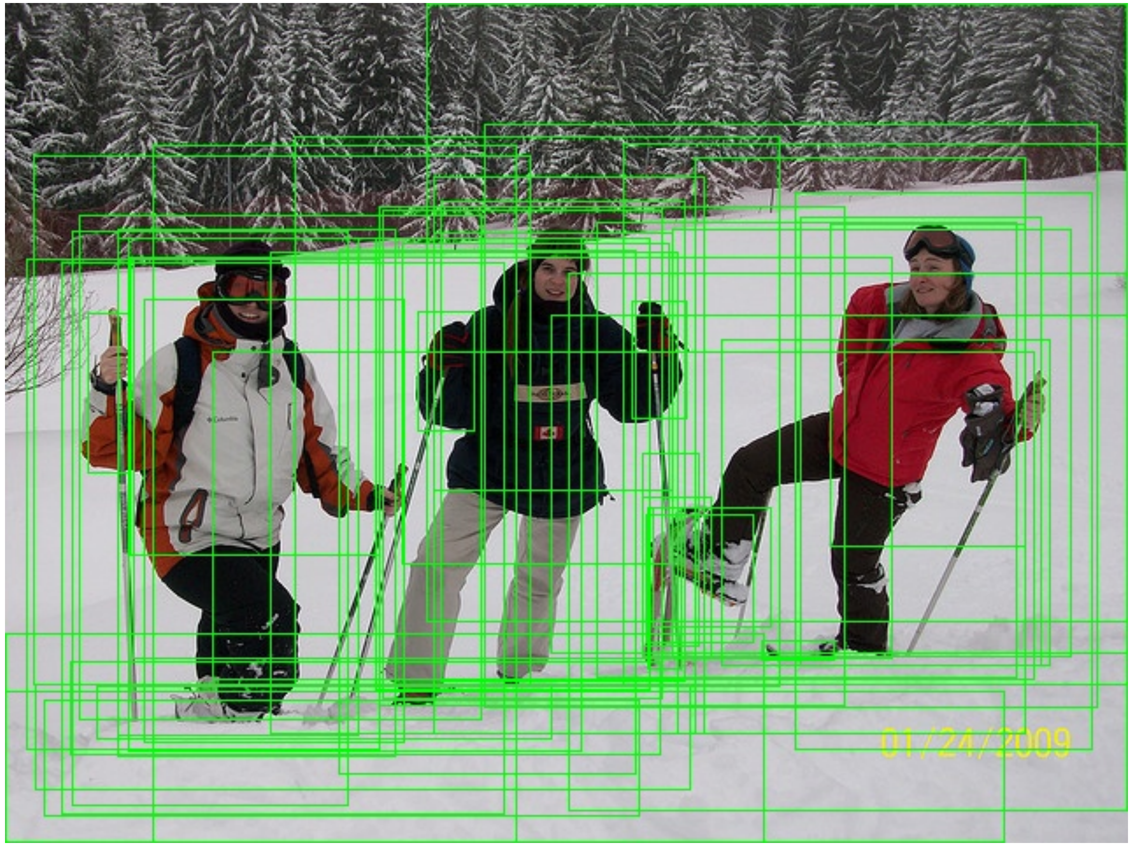
Method	Backbone	AR_{100}	AR_{300}	AR_{1000}	AR_S	AR_M	AR_L	Time
SharpMask [50]	ResNet-50	36.4	-	48.2	-	-	-	0.76
GCN-NS [42]	VGG-16 (Sync BN)	31.6	-	60.7	-	-	-	0.10
AttractioNet [21]	VGG-16	53.3	-	66.2	31.5	62.2	77.7	4.00
ZIP [32]	BN-inception	53.9	-	67.0	31.9	63.0	78.5	1.13
RPN [54]	ResNet-50-FPN	44.6	52.9	58.3	29.5	51.7	61.4	**0.04**
Iterative RPN		48.5	55.4	58.8	32.1	56.9	65.4	0.05
Iterative RPN+		54.0	60.4	63.0	35.6	62.7	73.9	0.06
GA-RPN [58]		59.1	65.1	68.5	40.7	68.2	78.4	0.06
Cascade RPN		**61.1**	**67.6**	**71.7**	**42.1**	**69.3**	**82.8**	0.06

Region proposal performance

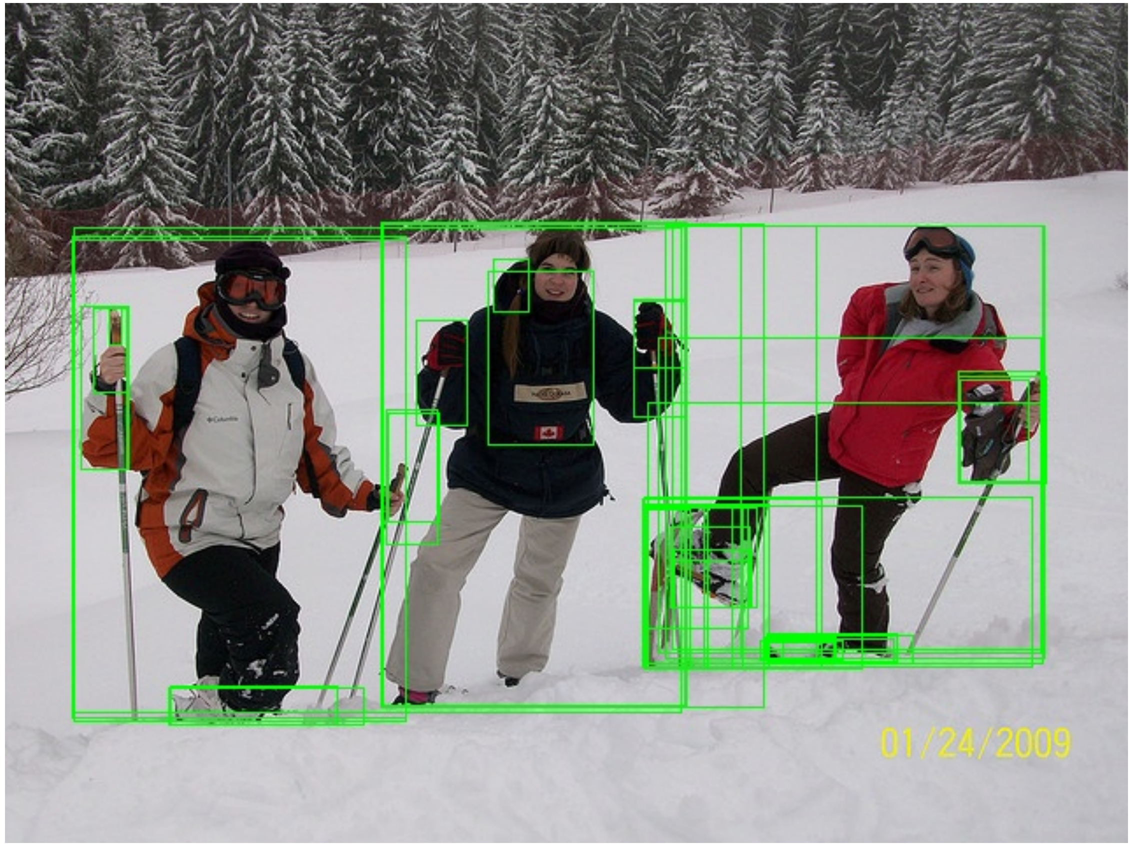
[50] Pinhero et al., ECCV 2016
[42] Lu et al., ECCV 2018
[21] Gidaris et al., arXiv 2016

[32] Li et al., IJCV 2019
[54] Ren et al., NeuIPS 2015
[58] Chen et al., CVPR 2019

Results

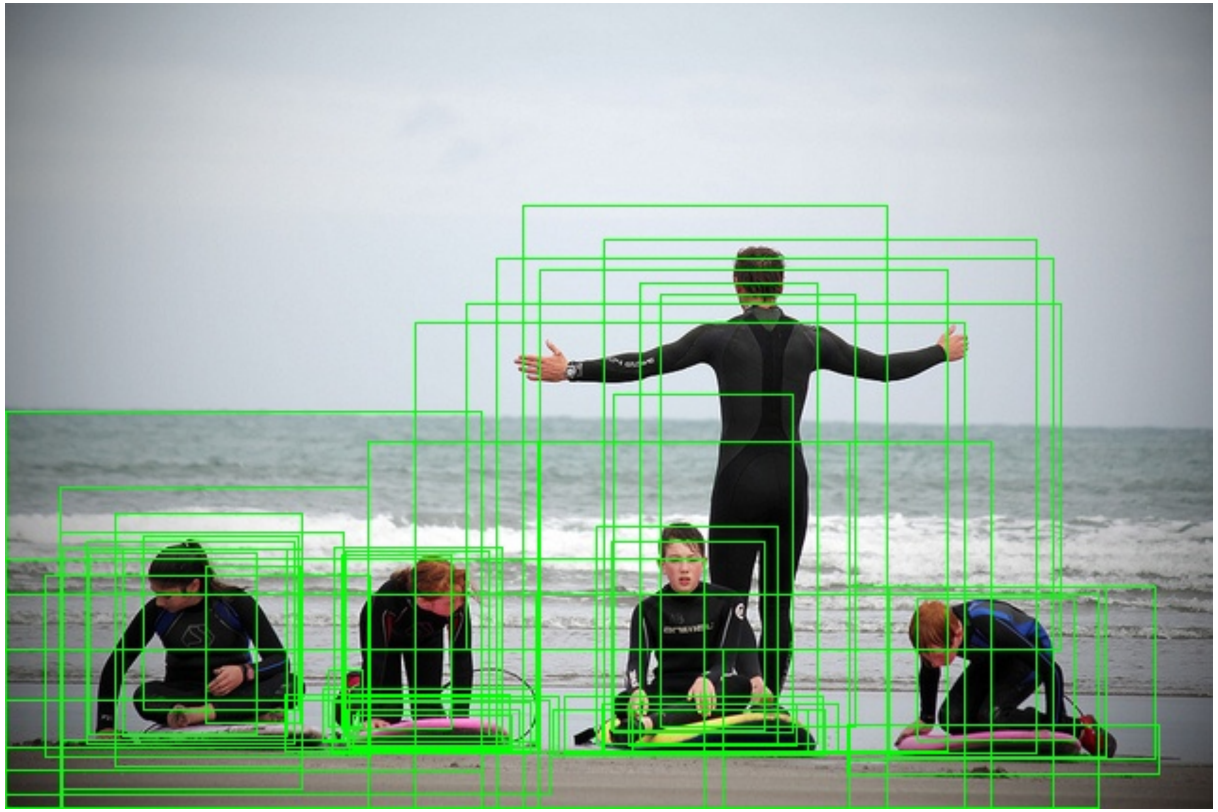


RPN

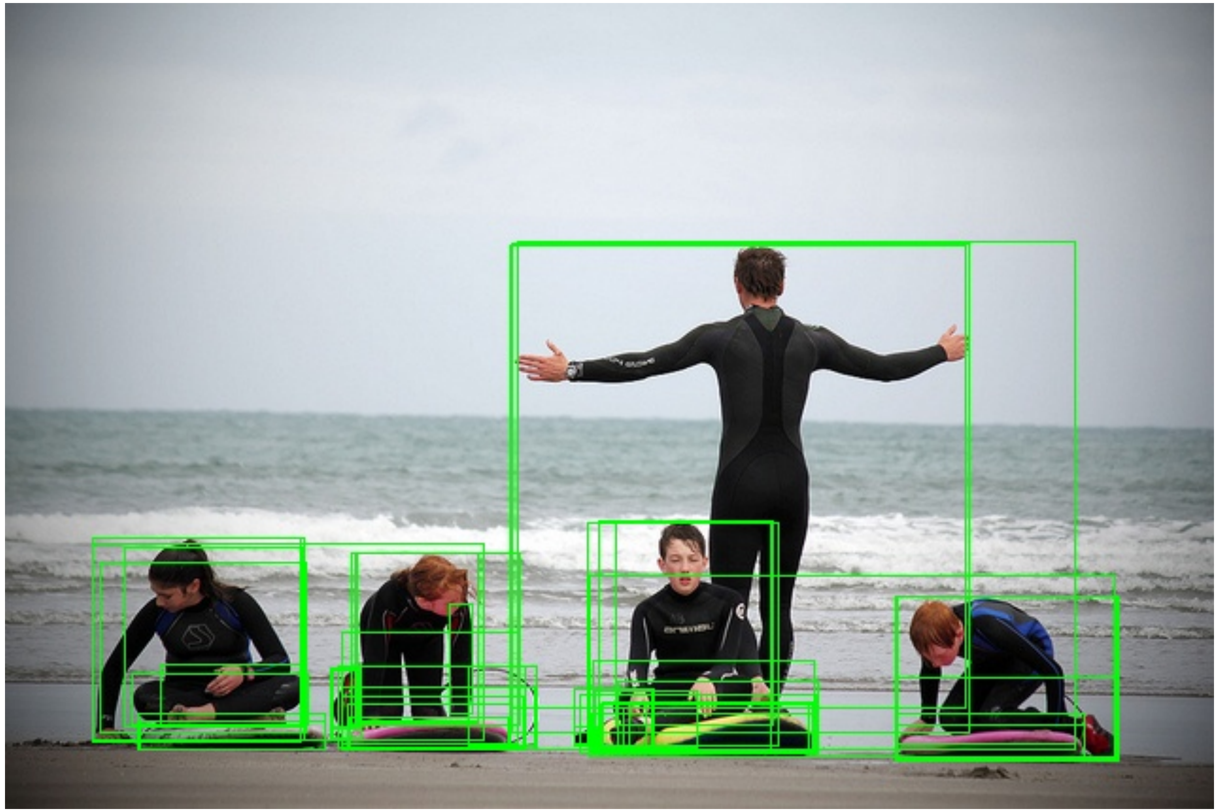


Cascade RPN

Results

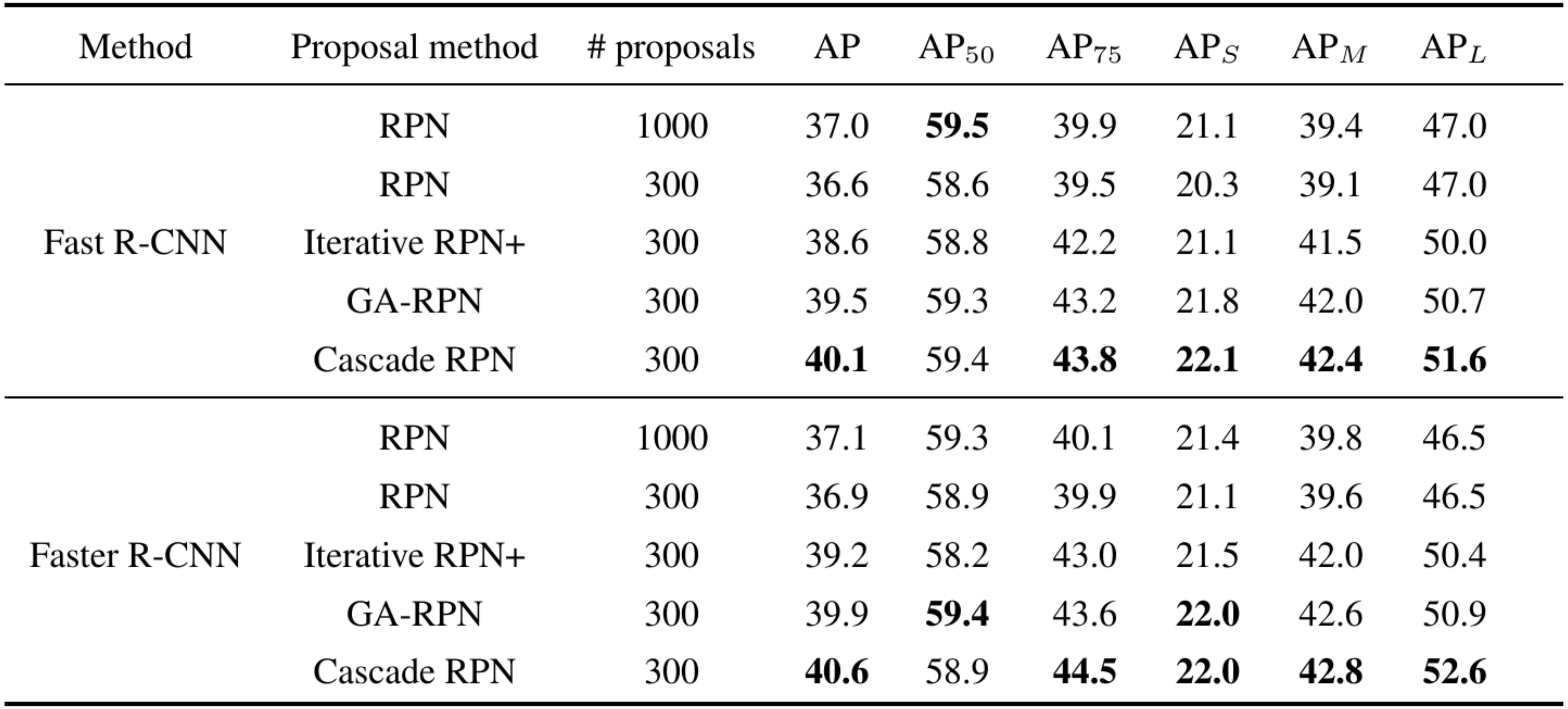


RPN



Cascade RPN

Results



Method	Proposal method	# proposals	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Fast R-CNN	RPN	1000	37.0	**59.5**	39.9	21.1	39.4	47.0
	RPN	300	36.6	58.6	39.5	20.3	39.1	47.0
	Iterative RPN+	300	38.6	58.8	42.2	21.1	41.5	50.0
	GA-RPN	300	39.5	59.3	43.2	21.8	42.0	50.7
	Cascade RPN	300	**40.1**	59.4	**43.8**	**22.1**	**42.4**	**51.6**
Faster R-CNN	RPN	1000	37.1	59.3	40.1	21.4	39.8	46.5
	RPN	300	36.9	58.9	39.9	21.1	39.6	46.5
	Iterative RPN+	300	39.2	58.2	43.0	21.5	42.0	50.4
	GA-RPN	300	39.9	**59.4**	43.6	**22.0**	42.6	50.9
	Cascade RPN	300	**40.6**	58.9	**44.5**	**22.0**	**42.8**	**52.6**

Detection performance when using different proposal methods

Conclusion

- Propose Cascade RPN for Object Detection
  - 13.4% higher recall than conventional RPN
  - Systematically maintain alignment between features and reference anchors

KAIST

Thank you!