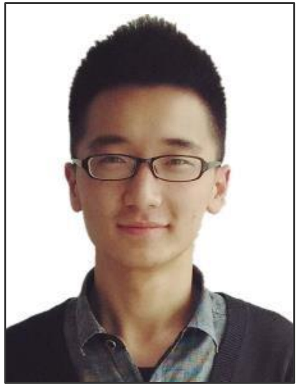
Bridging the Gap Between Anchor-based and Anchor-free Detection via Adaptive Training Sample Selection

Shifeng Zhang

CBSR, NLPR, CASIA

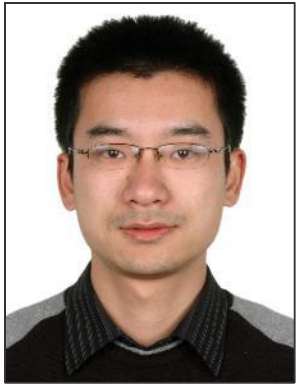
Authors



<u>**Shifeng Zhang**</u>

Cheng Chi

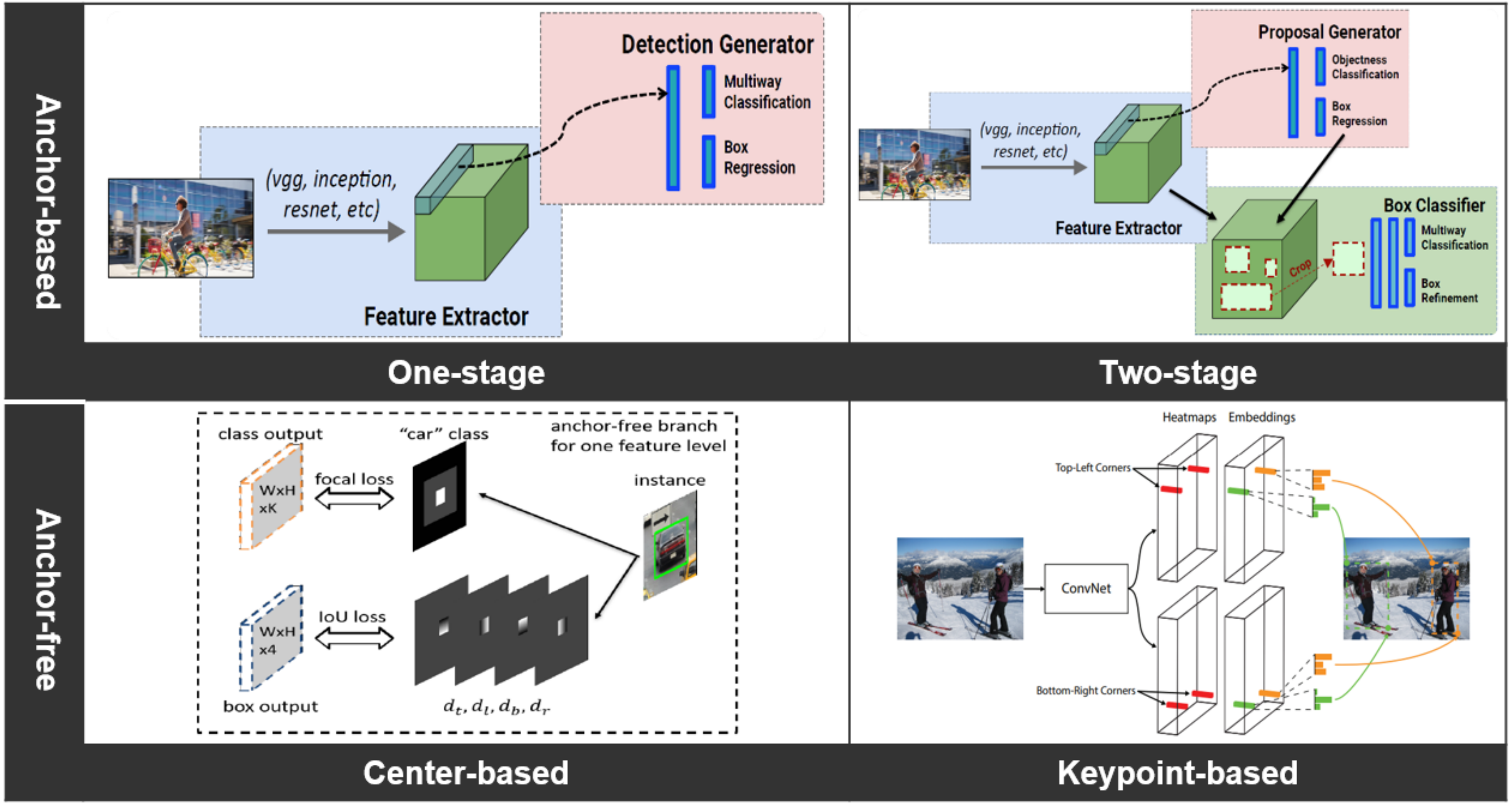
Yongqiang Yao



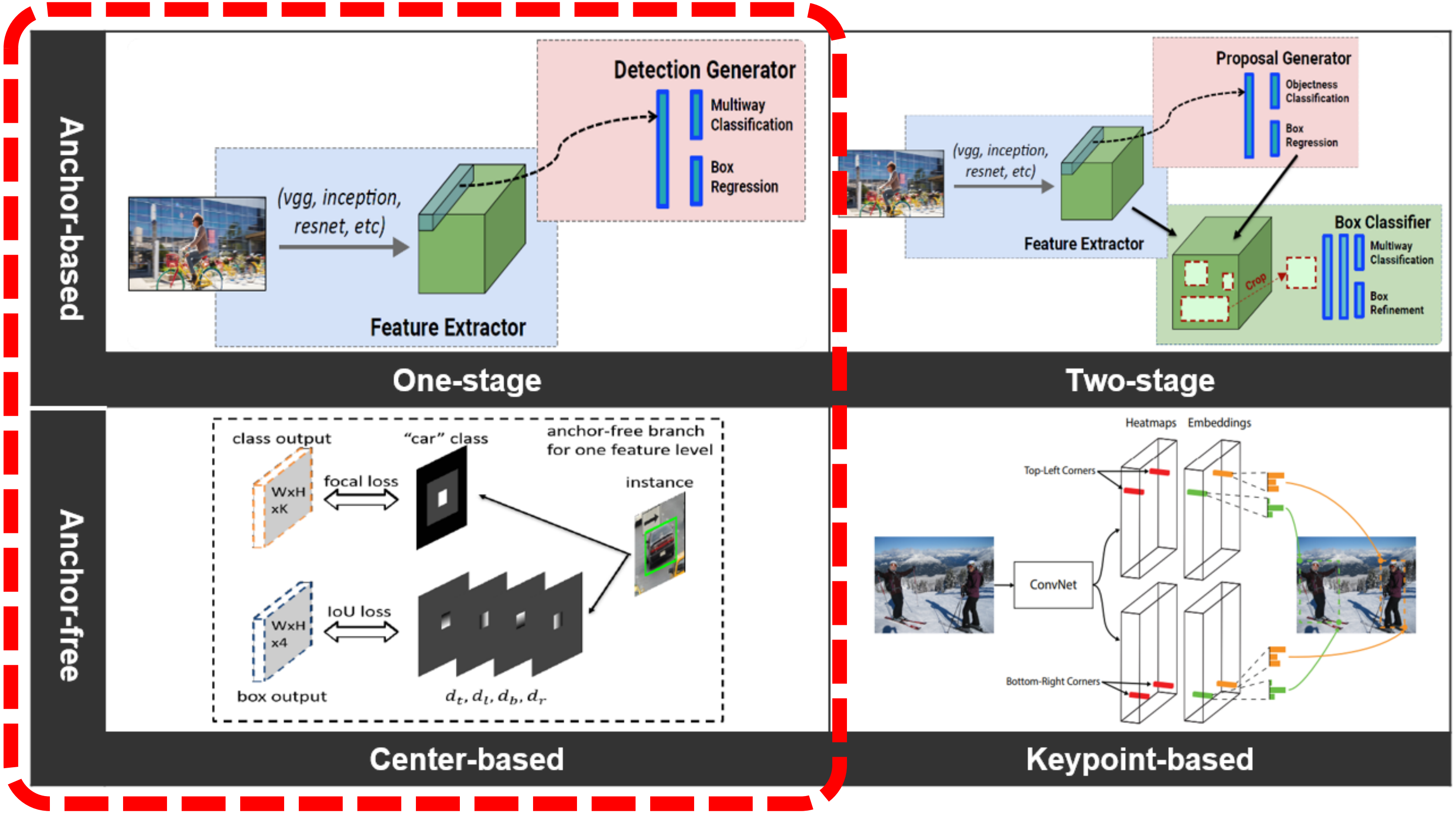
Zhen Lei

Stan Z. Li

Motivation


Anchor-based
Detection Generator
Multiway Classification
Box Regression
(vgg, inception, resnet, etc)
Feature Extractor
One-stage
Proposal Generator
Objectness Classification
Box Regression
Box Classifier
Multiway Classification
Box Refinement
Crop
Two-stage
Anchor-free
class output
"car" class
anchor-free branch for one feature level
focal loss
WxH xK
instance
IoU loss
WxH x4
box output
d_t, d_l, d_b, d_r
Center-based
Heatmaps
Embeddings
Top-Left Corners
ConvNet
Bottom-Right Corners
Keypoint-based

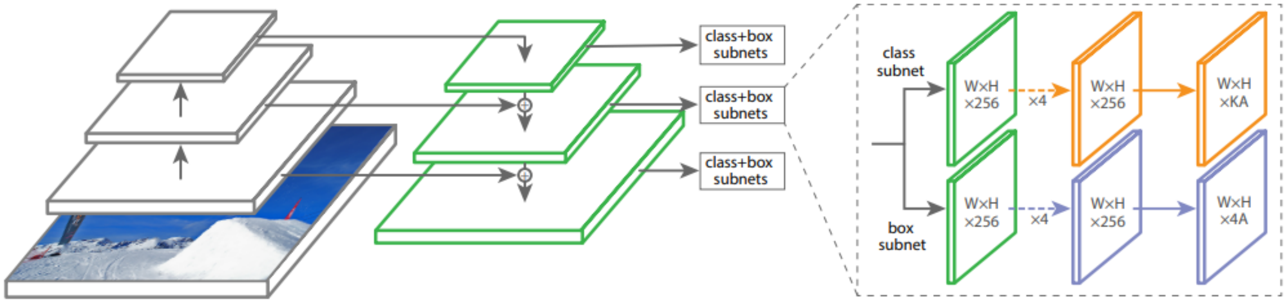
Motivation



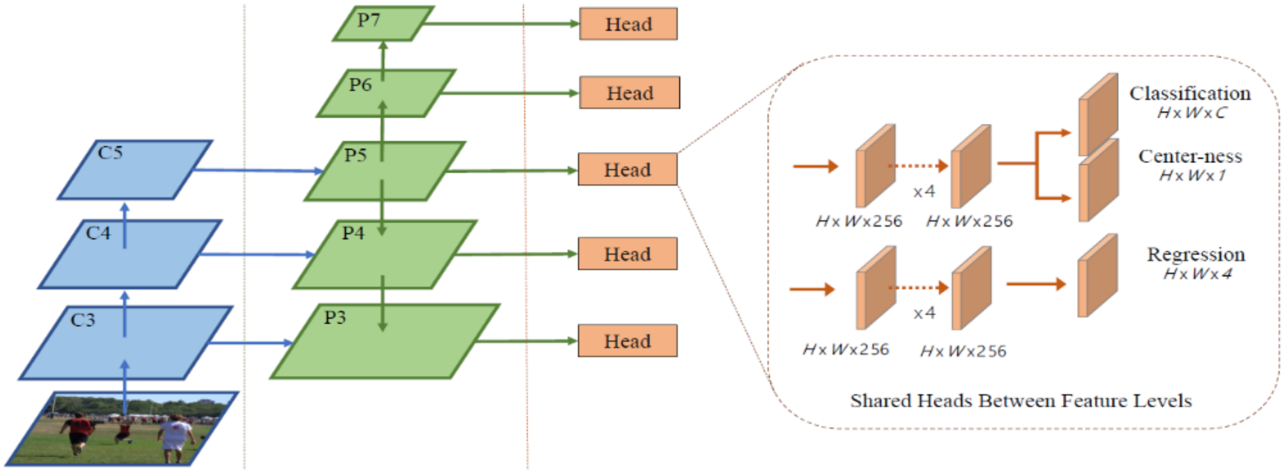
Anchor-based
Detection Generator
Multiway Classification
Box Regression
(vgg, inception, resnet, etc)
Feature Extractor
One-stage
Proposal Generator
Objectness Classification
Box Regression
(vgg, inception, resnet, etc)
Feature Extractor
Box Classifier
Multiway Classification
Box Refinement
Crop
Two-stage
Anchor-free
class output
"car" class
anchor-free branch for one feature level
WxH xK
focal loss
instance
WxH x4
IoU loss
box output
d_t, d_l, d_b, d_r
Center-based
Heatmaps
Embeddings
Top-Left Corners
ConvNet
Bottom-Right Corners
Keypoint-based

Motivation

Anchor-Based
RetinaNet



Anchor-free
FCOS

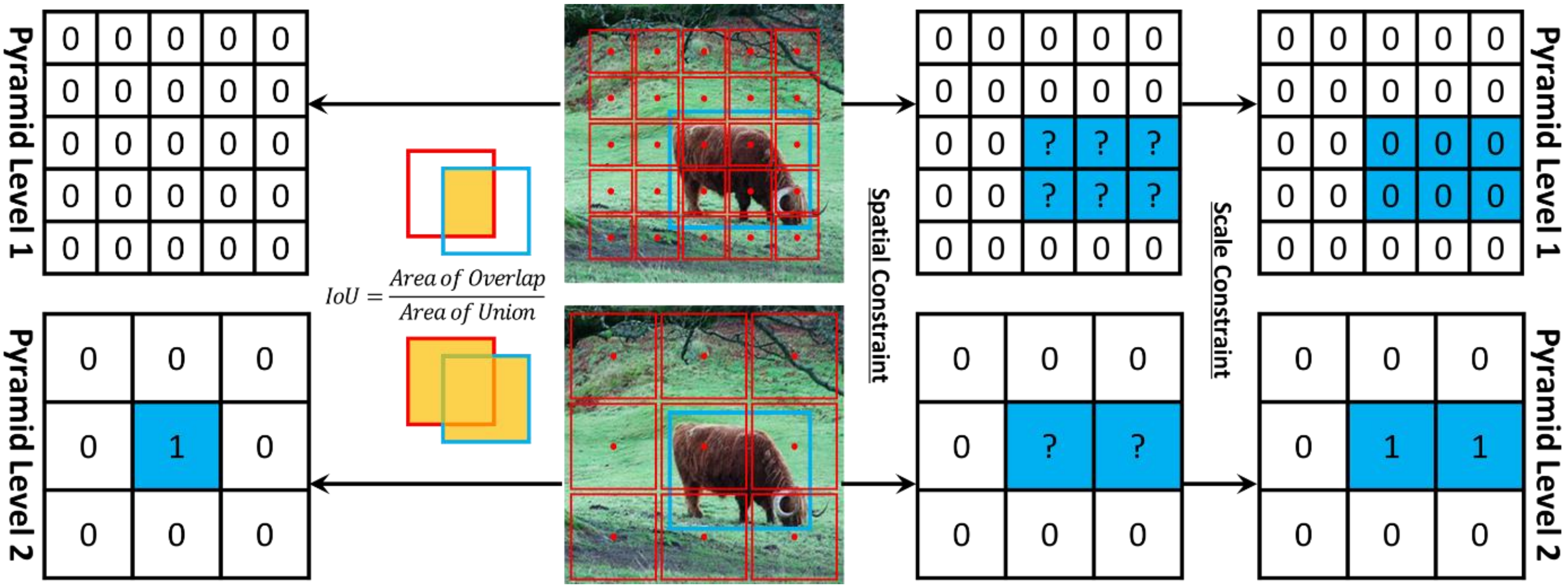


- Backbone
- FPN
- Cls head
- Reg head

Differences between RetinaNet and FCOS

Differences between RetinaNet and FCOS

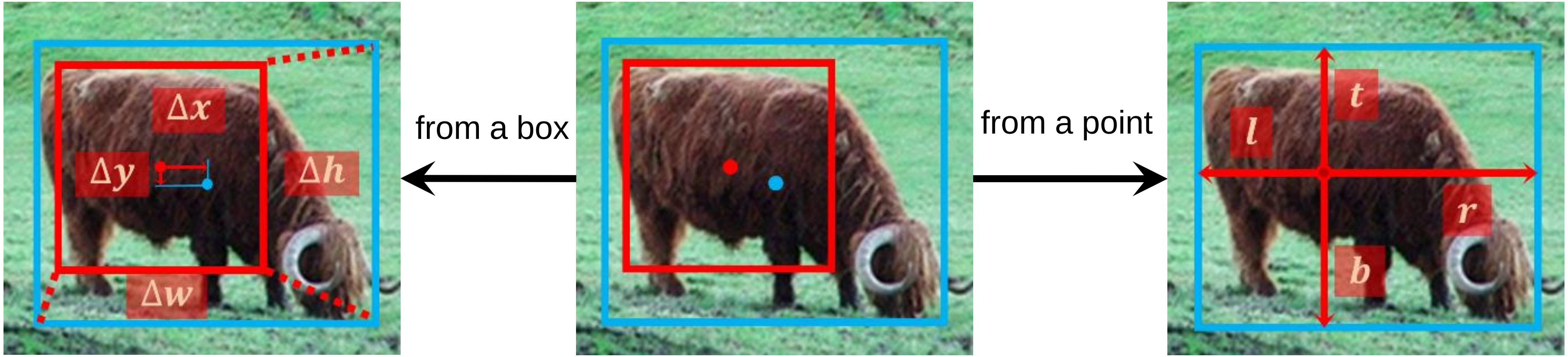
① Definition of positive and negative training samples



Differences between RetinaNet and FCOS

① Definition of positive and negative training samples

② Regression starting status

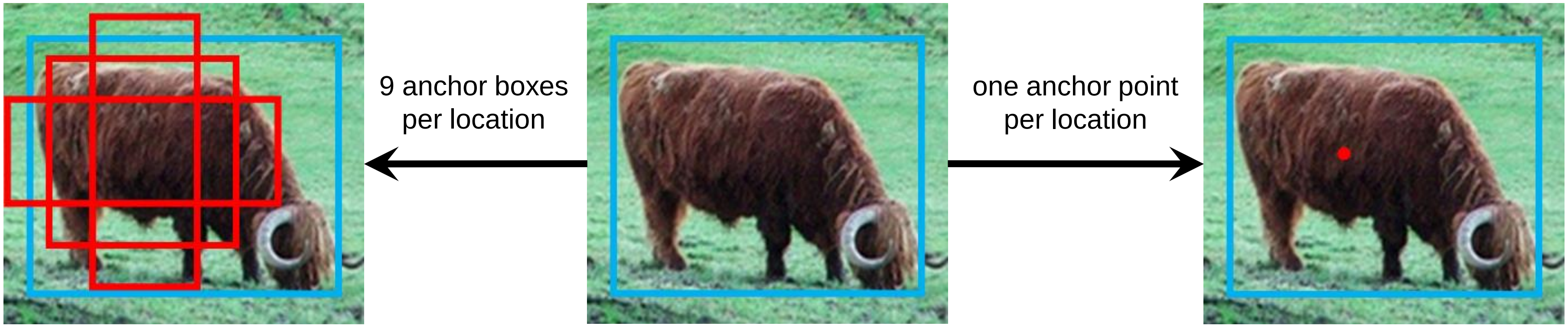


Differences between RetinaNet and FCOS

① Definition of positive and negative training samples

② Regression starting status

③ Number of anchors tiled per location



Differences between RetinaNet and FCOS

① Definition of positive and negative training samples

② Regression starting status

③ Number of anchors tiled per location

Which one is the essential difference?

Inconsistency Removal

① Definition of positive and negative training samples

② Regression starting status

Inconsistency	FCOS	RetinaNet (#A=1)					
GroupNorm	✓		✓	✓	✓	✓	✓
GIoU Loss	✓			✓	✓	✓	✓
In GT Box	✓				✓	✓	✓
Centerness	✓					✓	✓
Scalar	✓						✓
AP (%)	37.8	32.5	33.4	34.9	35.3	36.8	37.0

Inconsistency Removal

① Definition of positive and negative training samples

② Regression starting status

Inconsistency	FCOS	RetinaNet (#A=1)					
GroupNorm	✓		✓	✓	✓	✓	✓
GIoU Loss	✓			✓	✓	✓	✓
In GT Box	✓				✓	✓	✓
Centerness	✓					✓	✓
Scalar	✓						✓
AP (%)	37.8	32.5	33.4	34.9	35.3	36.8	37.0

Inconsistency Removal

① Definition of positive and negative training samples

② Regression starting status

Inconsistency	FCOS	RetinaNet (#A=1)					
GroupNorm	✓		✓	✓	✓	✓	✓
GIoU Loss	✓			✓	✓	✓	✓
In GT Box	✓				✓	✓	✓
Centerness	✓					✓	✓
Scalar	✓						✓
AP (%)	37.8	32.5	33.4	34.9	35.3	36.8	37.0

Inconsistency Removal

① Definition of positive and negative training samples

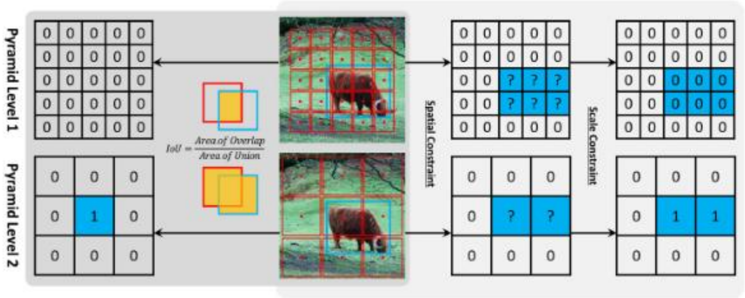
② Regression starting status

Inconsistency	FCOS	RetinaNet (#A=1)					
GroupNorm	✓		✓	✓	✓	✓	✓
GIoU Loss	✓			✓	✓	✓	✓
In GT Box	✓				✓	✓	✓
Centerness	✓					✓	✓
Scalar	✓						✓
AP (%)	37.8	32.5	33.4	34.9	35.3	36.8	37.0

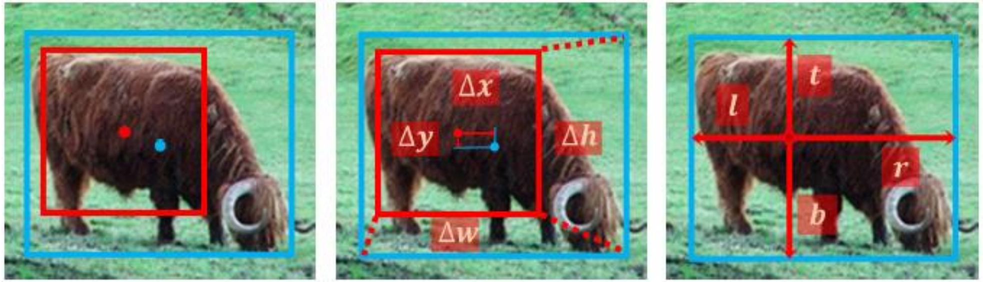
Essential Difference

① Definition of positive and negative training samples

② Regression starting status



(a) RetinaNet (b) FCOS



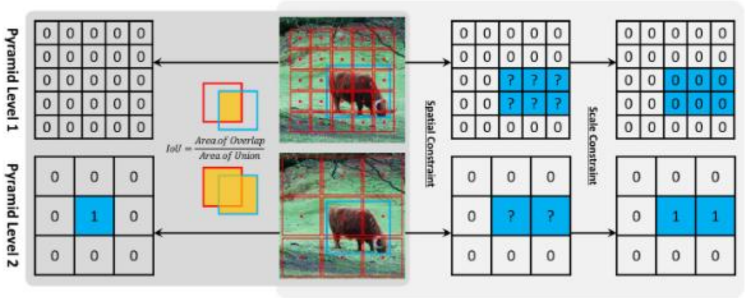
(a) Positive sample (b) RetinaNet (c) FCOS

Classification \ Regression	Box	Point
Intersection over Union	37.0	36.9
Spatial and Scale Constraint	37.8	37.8

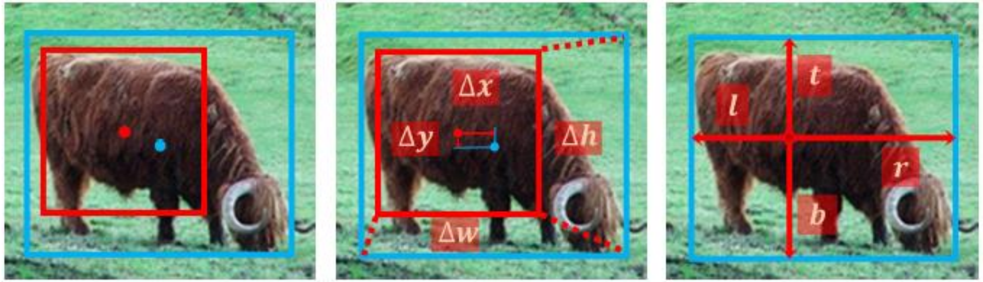
Essential Difference

① Definition of positive and negative training samples

② Regression starting status



(a) RetinaNet (b) FCOS



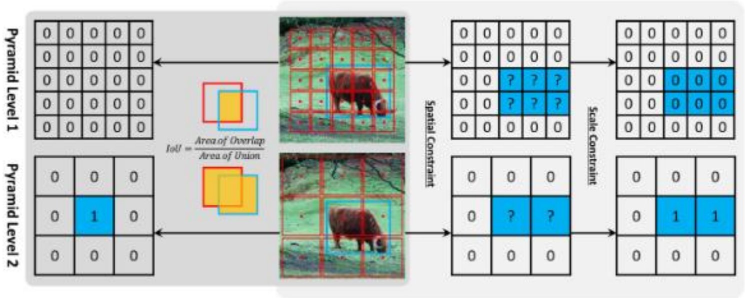
(a) Positive sample (b) RetinaNet (c) FCOS

Classification \ Regression	Box	Point
Intersection over Union	37.0	36.9
Spatial and Scale Constraint	37.8	37.8

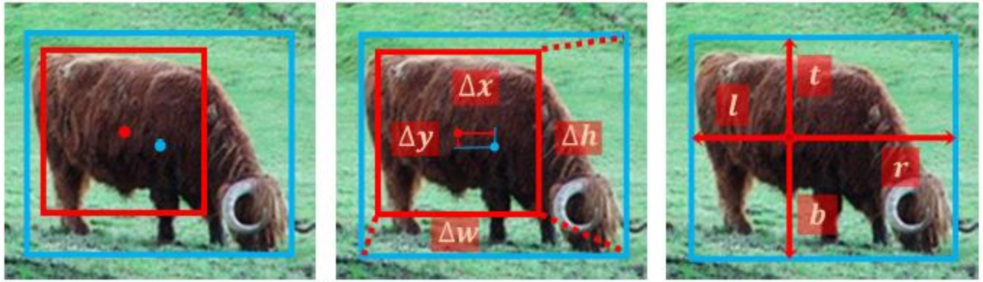
Essential Difference

① **Definition of positive and negative training samples**

~~② Regression starting status~~



(a) RetinaNet (b) FCOS



(a) Positive sample (b) RetinaNet (c) FCOS

Classification \ Regression	Box	Point
Intersection over Union	37.0	36.9
Spatial and Scale Constraint	37.8	37.8

Adaptive Training Sample Selection (ATSS)

① **Definition of positive and negative training samples**

~~② Regression starting status~~

Adaptive Training Sample Selection (ATSS)

① **Definition of positive and negative training samples**

~~② Regression starting status~~

$\boxed{\begin{array}{l}\textbf{for } \text{each level } i \in [1, \mathcal{L}] \textbf{ do} \\ \quad \mathcal{S}_i \leftarrow \text{ select } k \text{ anchors from } A_i \text{ whose center are closest} \\ \quad \text{to the center of ground-truth } g \text{ based on L2 distance;} \\ \quad \mathcal{C}_g = \mathcal{C}_g \cup \mathcal{S}_i; \\ \textbf{end for}\end{array}}$

compute IoU between $\mathcal{C}_g$ and g: $\mathcal{D}_g = IoU(\mathcal{C}_g, g)$;
compute mean of $\mathcal{D}_g$: $m_g = Mean(\mathcal{D}_g)$;
compute standard deviation of $\mathcal{D}_g$: $v_g = Std(\mathcal{D}_g)$;
compute IoU threshold for ground-truth g: $t_g = m_g + v_g$;

Adaptive Training Sample Selection (ATSS)

① **Definition of positive and negative training samples**

~~② Regression starting status~~

for each level $i \in [1, \mathcal{L}]$ **do**
  $\mathcal{S}_i \leftarrow$ select k anchors from A_i whose center are closest to the center of ground-truth g based on L2 distance;
  $\mathcal{C}_g = \mathcal{C}_g \cup \mathcal{S}_i$;
end for
compute IoU between $\mathcal{C}_g$ and g: $\mathcal{D}_g = IoU(\mathcal{C}_g, g)$;
compute mean of $\mathcal{D}_g$: $m_g = Mean(\mathcal{D}_g)$;
compute standard deviation of $\mathcal{D}_g$: $v_g = Std(\mathcal{D}_g)$;
compute IoU threshold for ground-truth g: $t_g = m_g + v_g$;

Adaptive Training Sample Selection (ATSS)

① **Definition of positive and negative training samples**

~~② Regression starting status~~

for each level $i \in [1, \mathcal{L}]$ **do**

$\mathcal{S}_i \leftarrow$ select k anchors from A_i whose center are closest to the center of ground-truth g based on L2 distance;

$\mathcal{C}_g = \mathcal{C}_g \cup \mathcal{S}_i$;

end for

compute IoU between $\mathcal{C}_g$ and g: $\mathcal{D}_g = IoU(\mathcal{C}_g, g)$;

compute mean of $\mathcal{D}_g$: $m_g = Mean(\mathcal{D}_g)$;

compute standard deviation of $\mathcal{D}_g$: $v_g = Std(\mathcal{D}_g)$;

compute IoU threshold for ground-truth g: $t_g = m_g + v_g$;

Adaptive Training Sample Selection (ATSS)

① **Definition of positive and negative training samples**

~~② Regression starting status~~

for each level $i \in [1, \mathcal{L}]$ **do**
  $\mathcal{S}_i \leftarrow$ select k anchors from A_i whose center are closest to the center of ground-truth g based on L2 distance;
  $\mathcal{C}_g = \mathcal{C}_g \cup \mathcal{S}_i$;
end for
compute IoU between $\mathcal{C}_g$ and g: $\mathcal{D}_g = IoU(\mathcal{C}_g, g)$;
compute mean of $\mathcal{D}_g$: $m_g = Mean(\mathcal{D}_g)$;
compute standard deviation of $\mathcal{D}_g$: $v_g = Std(\mathcal{D}_g)$;
$\boxed{\text{compute IoU threshold for ground-truth } g\text{: } t_g = m_g + v_g;}$

Adaptive Training Sample Selection (ATSS)

① **Definition of positive and negative training samples**

~~② Regression starting status~~

for each level $i \in [1, \mathcal{L}]$ **do**

$\mathcal{S}_i \leftarrow$ select k anchors from A_i whose center are closest to the center of ground-truth g based on L2 distance;

$\mathcal{C}_g = \mathcal{C}_g \cup \mathcal{S}_i$;

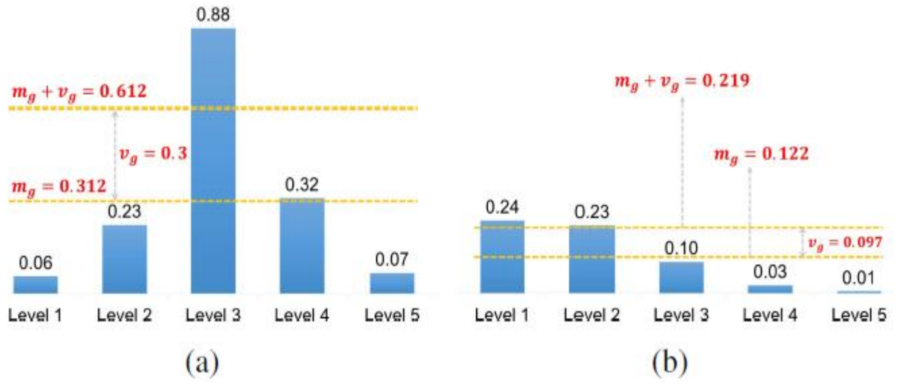
end for

compute IoU between $\mathcal{C}_g$ and g: $\mathcal{D}_g = IoU(\mathcal{C}_g, g)$;

compute mean of $\mathcal{D}_g$: $m_g = Mean(\mathcal{D}_g)$;

compute standard deviation of $\mathcal{D}_g$: $v_g = Std(\mathcal{D}_g)$;

compute IoU threshold for ground-truth g: $t_g = m_g + v_g$;

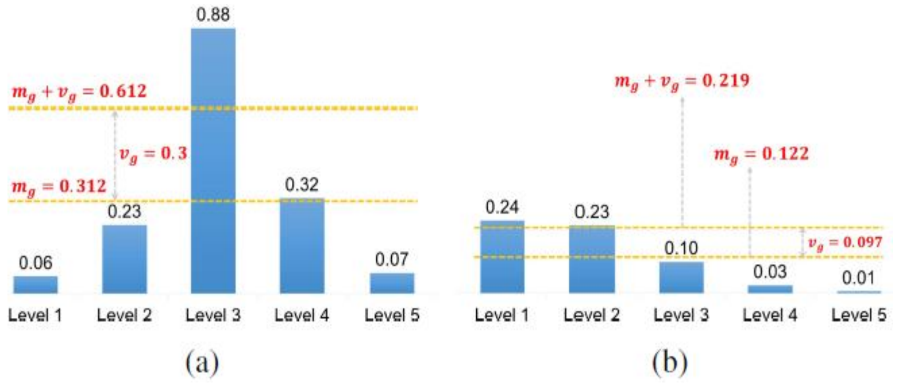


Adaptive Training Sample Selection (ATSS)

① **Definition of positive and negative training samples**

~~② Regression starting status~~

for each level $i \in [1, \mathcal{L}]$ **do**
  $\mathcal{S}_i \leftarrow$ select k anchors from A_i whose center are closest to the center of ground-truth g based on L2 distance;
  $\mathcal{C}_g = \mathcal{C}_g \cup \mathcal{S}_i$;
end for
compute IoU between $\mathcal{C}_g$ and g: $\mathcal{D}_g = IoU(\mathcal{C}_g, g)$;
compute mean of $\mathcal{D}_g$: $m_g = Mean(\mathcal{D}_g)$;
compute standard deviation of $\mathcal{D}_g$: $v_g = Std(\mathcal{D}_g)$;
compute IoU threshold for ground-truth g: $t_g = m_g + v_g$;



Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet (#A=1)	37.0	55.1	39.9	21.4	41.2	48.6
RetinaNet (#A=1) + ATSS	39.3	57.5	42.8	24.3	43.3	51.3
FCOS	37.8	55.6	40.7	22.1	41.8	48.8
FCOS + Center sampling	38.6	57.4	41.4	22.3	42.5	49.8
FCOS + ATSS	39.2	57.3	42.4	22.7	43.1	51.5

ResNet-101	43.6	62.1	47.4	26.1	47.0	53.6
ResNeXt-32x8d-101	45.1	63.9	49.1	27.9	48.2	54.6
ResNeXt-64x4d-101	45.6	64.6	49.7	28.5	48.9	55.6
ResNet-101-DCN	46.3	64.7	50.4	27.7	49.8	58.4
ResNeXt-32x8d-101-DCN	47.7	66.6	52.1	29.3	50.8	59.7
ResNeXt-64x4d-101-DCN	47.7	66.5	51.9	29.7	50.8	59.4
ResNeXt-32x8d-101-DCN	50.6	68.6	56.1	**33.6**	**52.9**	62.2
ResNeXt-64x4d-101-DCN	**50.7**	**68.9**	**56.3**	33.2	**52.9**	**62.4**

Analysis

k	3	5	7	9	11	13	15	17	19
AP (%)	38.0	38.8	39.1	39.3	39.1	39.0	39.1	39.2	38.9

- Hyperparameter k is quite robust

Analysis

k	3	5	7	9	11	13	15	17	19
AP (%)	38.0	38.8	39.1	39.3	39.1	39.0	39.1	39.2	38.9

Scale	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
5	39.0	57.9	41.9	23.2	42.8	50.5
6	39.2	57.6	42.5	23.5	42.8	51.1
7	39.3	57.6	42.4	22.9	43.2	51.3
8	39.3	57.5	42.8	24.3	43.3	51.3
9	38.9	56.5	42.0	22.9	42.4	50.3

Aspect Ratio	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
4:1	39.1	57.2	42.3	23.1	43.1	51.4
2:1	39.0	56.9	42.5	23.3	43.5	50.6
1:1	39.3	57.5	42.8	24.3	43.3	51.3
2:1	39.3	57.4	42.3	22.8	43.4	51.0
4:1	39.1	56.9	42.6	22.9	42.9	50.7

- Hyperparameter k is quite robust
- ATSS only tiles one anchor boxes per location
- ATSS is robust to different anchor scales
- ATSS is robust to different anchor aspect ratios

Discussion

① Definition of positive and negative training samples

② ~~Regression starting status~~

③ **Number of anchors tiled per location?**

Inconsistency	FCOS	RetinaNet (#A=1)					
GroupNorm	✓		✓	✓	✓	✓	✓
GIoU Loss	✓			✓	✓	✓	✓
In GT Box	✓				✓	✓	✓
Centerness	✓					✓	✓
Scalar	✓						✓
AP (%)	37.8	32.5	33.4	34.9	35.3	36.8	37.0

Method	#sc	#ar	AP	AP_{50}	AP_{75}
RetinaNet (#A=9)	3	3	36.3	55.2	38.8
+Imprs.	3	3	38.4	56.2	41.6
+Imprs.+ATSS	3	3	39.2	57.6	42.7
+Imprs.+ATSS	3	1	39.3	57.7	42.6
+Imprs.+ATSS	1	3	39.2	57.1	42.5
+Imprs.+ATSS	1	1	39.3	57.5	42.8

- Under the IoU-based sample selection strategy, tiling more anchor boxes per location is effective

Discussion

① Definition of positive and negative training samples

② ~~Regression starting status~~

③ **Number of anchors tiled per location?**

Inconsistency	FCOS	RetinaNet (#A=1)					
GroupNorm	✓		✓	✓	✓	✓	✓
GIoU Loss	✓			✓	✓	✓	✓
In GT Box	✓				✓	✓	✓
Centerness	✓					✓	✓
Scalar	✓						✓
AP (%)	37.8	32.5	33.4	34.9	35.3	36.8	37.0

Method	#sc	#ar	AP	AP_{50}	AP_{75}
RetinaNet (#A=9)	3	3	36.3	55.2	38.8
+Imprs.	3	3	38.4	56.2	41.6
+Imprs.+ATSS	3	3	39.2	57.6	42.7
+Imprs.+ATSS	3	1	39.3	57.7	42.6
+Imprs.+ATSS	1	3	39.2	57.1	42.5
+Imprs.+ATSS	1	1	39.3	57.5	42.8

- Under the IoU-based sample selection strategy, tiling more anchor boxes per location is effective
- Under the proposed ATSS, tiling multiple anchors per location is a useless operation

Discussion

① Definition of positive and negative training samples

~~② Regression starting status~~

③ Number of anchors tiled per location?

Inconsistency	FCOS	RetinaNet (#A=1)					
GroupNorm	✓		✓	✓	✓	✓	✓
GIoU Loss	✓			✓	✓	✓	✓
In GT Box	✓				✓	✓	✓
Centerness	✓					✓	✓
Scalar	✓						✓
AP (%)	37.8	32.5	33.4	34.9	35.3	36.8	37.0

Method	#sc	#ar	AP	AP_{50}	AP_{75}
RetinaNet (#A=9)	3	3	36.3	55.2	38.8
+Imprs.	3	3	38.4	56.2	41.6
+Imprs.+ATSS	3	3	39.2	57.6	42.7
+Imprs.+ATSS	3	1	39.3	57.7	42.6
+Imprs.+ATSS	1	3	39.2	57.1	42.5
+Imprs.+ATSS	1	1	39.3	57.5	42.8

- Under the IoU-based sample selection strategy, tiling more anchor boxes per location is effective
- Under the proposed ATSS, tiling multiple anchors per location is a useless operation
- Select positive samples appropriately, different number of anchors per location has the same result
- Further study is needed to discover the right role of tiling multiple anchors per location

Thanks