上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Bidirectional Matrix Feature Pyramid Network for Object Detection

饮 水 思 源 · 爱 国 荣 校

Wei XU, Yi Gan, Jianbo Su

SJTU

上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

目 录

Content

1 Introduction

2 Proposed methods

3 Experiments

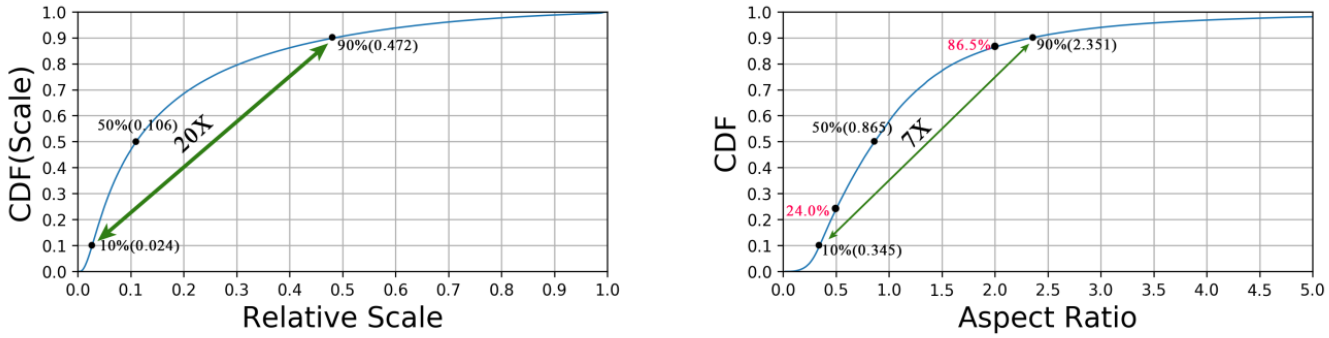
4 Conclusion

01

Introduction

Statistical analysis for MS COCO

Scale variation is one of the key challenges in object detection.



(a) Fraction of bboxes *vs* scale of bboxes relative to the image

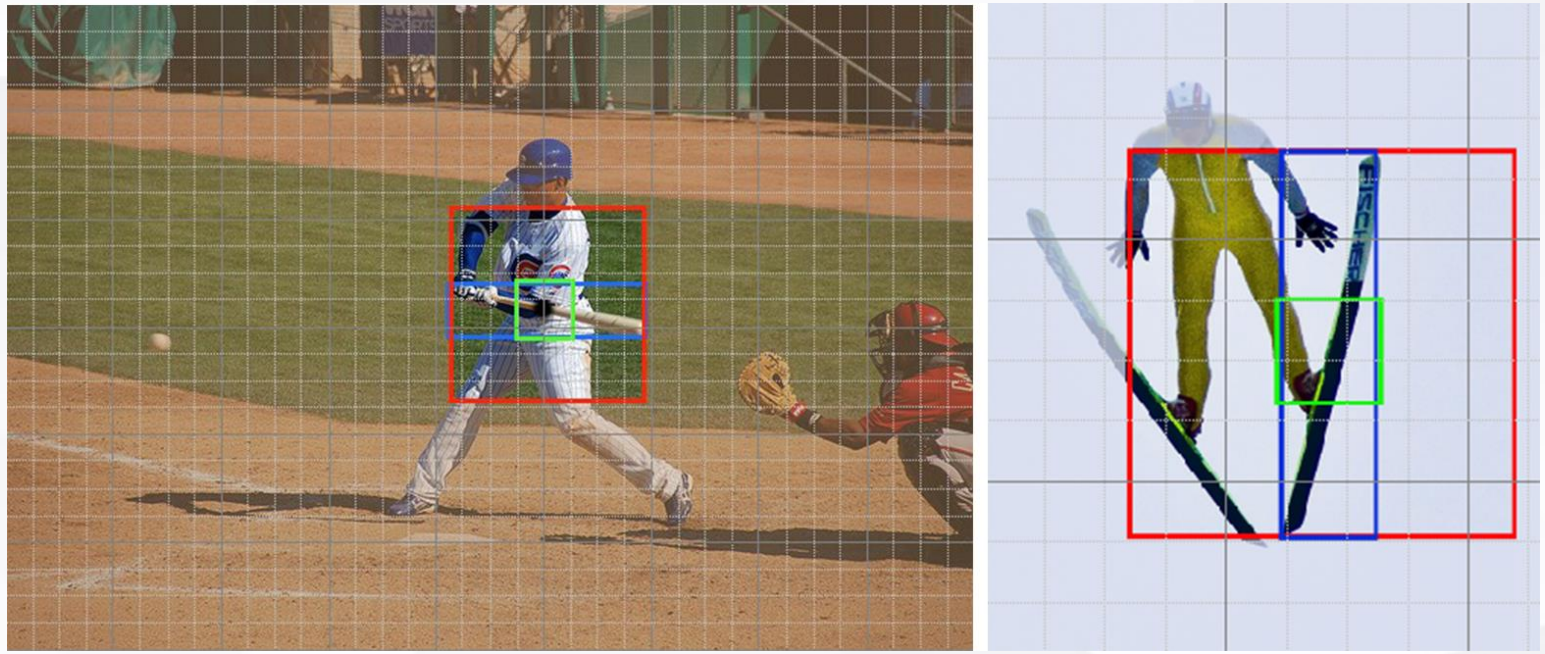
(b) Fraction of bboxes *vs* aspect ratio of bboxes

Fig. 1. Statistical analysis of annotated bounding boxes (bboxes) in COCO. CDF is the abbreviation of cumulative distribution function.

Not only the scale variation but also the aspect ratio variation should be taken into account.

Examples of objects assignment

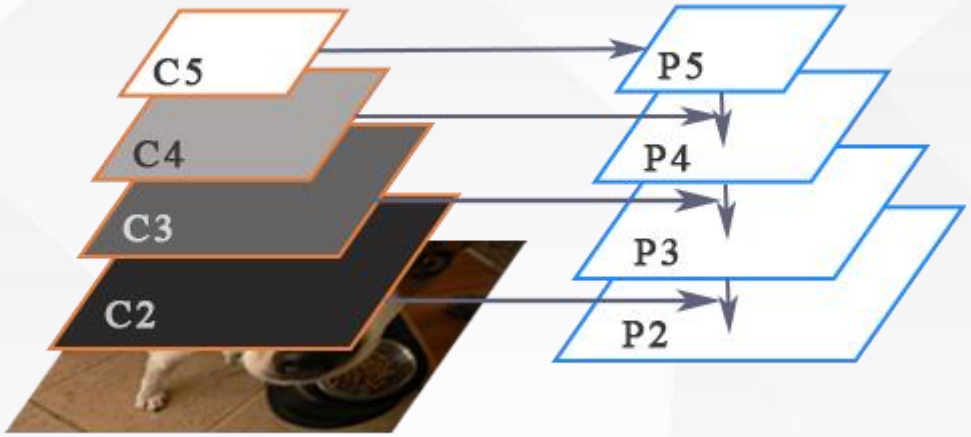
Feature pyramids are widely used to improve scale invariance by mapping the objects to feature maps with relevant square receptive fields.



The regions in the red and green rectangles are the receptive fields of feature layers at different resolutions respectively, when mapping the objects in the blue rectangle.

The poor match between objects and assigned features is bound to occur among a rectangular and a square receptive filed.

Feature fusion



FPN [1] only forms a single top-down pathway to propagate high-level information. It will make the integrated features focus more on adjacent resolution but less on others. Each feature in the pyramid may mainly or only contain single-level information, thus limiting the detection performance.

[1] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2117–2125, 2017.

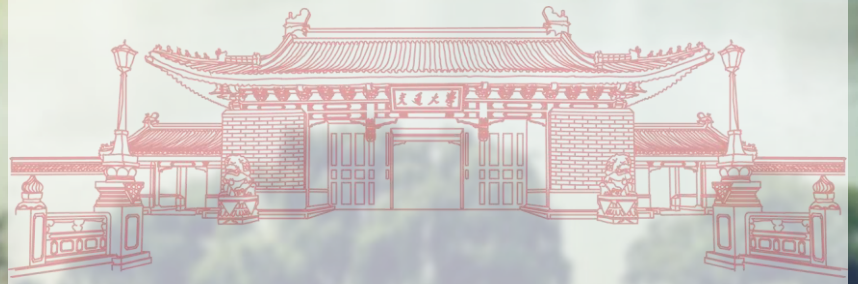
Motivations

The designed feature pyramids should match objects with multiple scales and aspect ratios. Specifically, objects should be assigned to feature layers with fitted receptive filed.

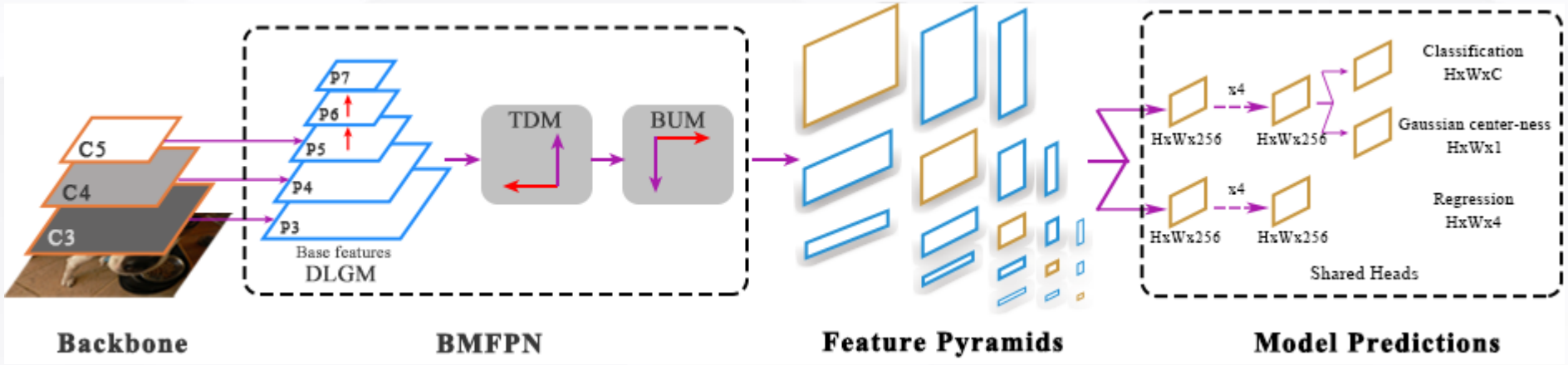
Each feature in the pyramid is required to be representative enough and contain rich information. Specifically, the feature fusion methods should combine low-level and high-level information effectively.

02

Proposed methods



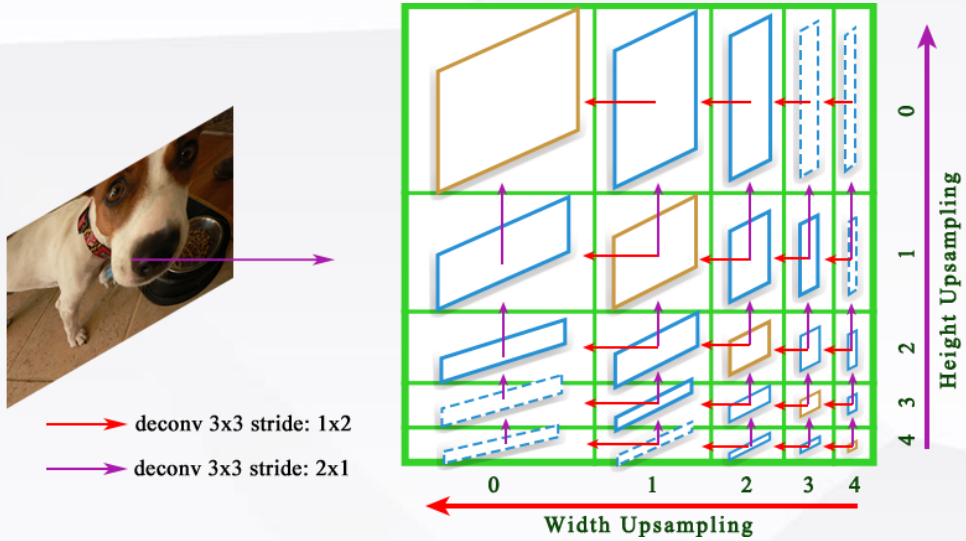
Overall pipeline



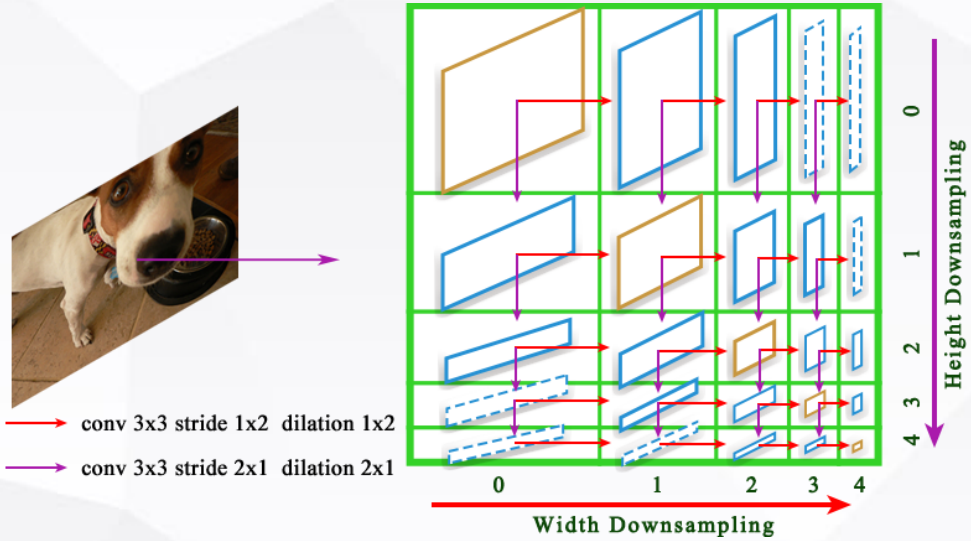
This is the overall pipeline of our model based on FCOS[1]. DLGM utilizes multi-level features extracted by backbone to generate the base features. Then the base features are fed into TDM and BUM in series to construct feature pyramids for final model predictions.

[1] Z. Tian, C. Shen, H. Chen, and T. He, “Fcos: Fully convolutional one-stage object detection,” *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9627–9636, 2019

BMFPN



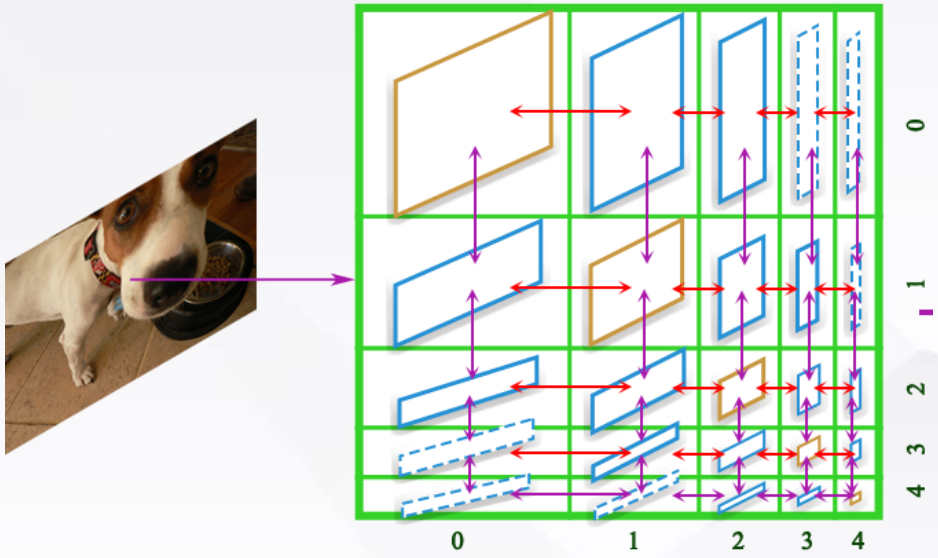
TDM



BUM

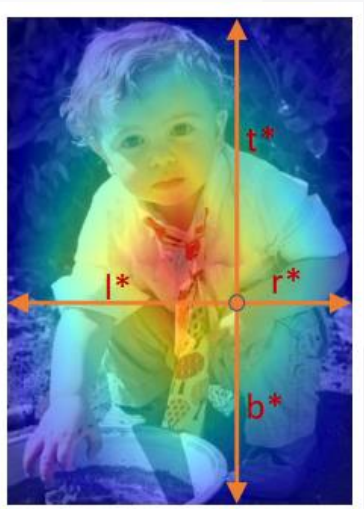
Deconvolution and convolution with asymmetric strides in h and w dimensions to construct the final feature pyramids.

Object assignment strategy



The scale and aspect ratio of receptive field are proportional between each feature layer in theory.

we set both the width (WR) and height ranges(HR) of object instances assigned to each feature layer.



For example:

$L_{(0,0)}$: WR: *(16, 32)* *HR:(16, 32)*

$L_{(0,1)}$: WR: *(16, 32)* *HR:(32, 64)*

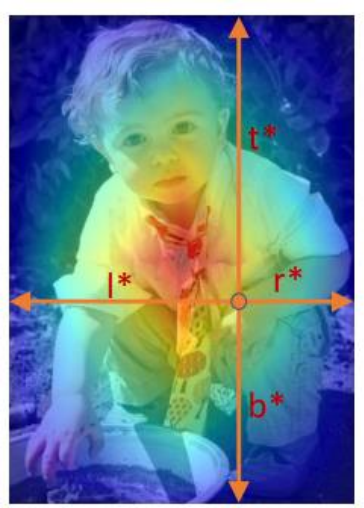
$L_{(1,0)}$: WR: *(32, 64)* *HR:(16, 32)*

if a location satisfies：

$$\max(l^*, r^*) \in WR_{L_{(i,j)}}, \max(b^*, t^*) \in HR_{L_{(i,j)}}$$

it will be assigned to layer $L_{(i,j)}$.

Gaussian Center-ness Branch



$$\text{Gauss\_centerness}^* = \sqrt{e^{-\frac{(x_i-x_0)^2}{2\sigma_x^2}-\frac{(y_i-y_0)^2}{2\sigma_y^2}}}$$

$$\sigma_x = \frac{w}{6\sigma}, \sigma_y = \frac{h}{6\sigma}$$

$$w = l^* + r^*, h = t^* + b^*$$

$$x_i - x_0 = \frac{|l^*-r^*|}{2}, y_i - y_0 = \frac{|t^*-b^*|}{2}$$

The target is decided by the hyper-parameter σ, central location (x_0, y_0) and box size (h, w) .

03

Experiments

Experiments

TABLE I

THE PERFORMANCE FROM THE BASELINE GRADUALLY TO ALL COMPONENTS INCORPORATED ON MS COCO VAL-2017 SPLIT.

GCB	DLGM	TDM	BUM	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
				36.7	55.8	39.2	21.0	40.7	48.4
✓				37.0	55.9	39.5	21.3	40.9	48.6
	✓			33.2	49.9	35.4	14.5	36.1	47.8
	✓	✓		37.6	56.7	40.3	21.0	41.5	49.6
	✓		✓	36.5	53.0	39.3	16.9	40.6	52.2
	✓	✓	✓	39.7	57.7	42.8	22.4	44.4	53.0
✓	✓	✓	✓	**40.0**	**57.7**	**43.1**	**22.8**	**44.5**	**53.4**

With all these components added to FCOS, improvement on AP is 3.3% over baseline. And the results shows that large size instances contribute most (+5.0%). Moreover, it makes more accurate detection with 3.9% improvement on AP_{75}.

Ablation Studies

TABLE III
COMPARISON BETWEEN DIFFERENT FEATURE PYRAMID STRUCTURES BASED ON FCOS.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params (M)
Baseline(FPN) [17]	36.7	55.8	39.2	21.0	40.7	48.4	32.02
PAFPN [21]	37.0	56.1	39.4	20.8	40.8	48.4	34.38
xNets [27]	37.9	56.0	40.7	21.1	43.0	50.9	33.20
BMFPN*	38.0	56.3	40.5	21.2	41.7	51.3	34.97
BMFPN	**39.7**	**57.7**	**42.8**	**22.4**	**44.4**	**53.0**	34.97

TABLE IV
COMPARISON BETWEEN DIFFERENT CONNECTION MODE ON MS COCO VAL-2017 SPLIT. TDM+BUM : PARALLEL MODE; BUM→TDM : BUM FIRST UNDER SERIES MODE; TDM→BUM : TDM FIRST UNDER SERIES MODE.

Mode	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
TDM+BUM	38.9	57.5	41.6	22.2	43.2	51.8
BUM→TDM	39.2	57.6	42.1	22.0	43.5	51.9
TDM→BUM	**39.7**	**57.7**	**42.8**	**22.4**	**44.4**	**53.0**

TABLE V
COMPARISON WITH DIFFERENT CONNECTION LAYERS BETWEEN TDM AND BUM ON MS COCO VAL-2017 SPLIT. MODE NO DENOTES TDM AND BUM ARE DIRECTLY CONNECTED WITHOUT ANY OTHER CONNECTION LAYER.

Mode	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
NO	39.4	57.7	42.1	22.3	43.6	53.3
Convolution	39.4	57.6	42.2	21.5	44.2	52.9
Dilated convolution	**39.7**	**57.7**	**42.8**	**22.4**	**44.1**	**53.0**

Experiments

TABLE VI
COMPARISON WITH STATE-OF-THE-ART DETECTORS ON MS COCO TEST-DEV SPLIT.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage detectors							
Faster R-CNN w/FPN [17]	ResNet-101	36.2	59.1	39.0	18.2	39.0	48.2
Mask R-CNN [33]	ResNeXt-101	39.8	62.3	43.4	22.1	43.2	51.2
Cascade R-CNN [34]	ResNet-101	42.8	62.1	46.3	23.7	45.5	55.2
D-RFCN + SNIP [16]	DPN-98	45.7	67.3	51.1	29.3	48.8	57.1
TridentNet [35]	ResNet-101-DCN	46.8	67.6	51.5	28.0	51.2	60.5
One-stage detectors							
YOLOv3-608 [36]	DarkNet-53	33.0	57.9	34.4	18.3	35.4	41.9
SSD513 [7]	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8
RetinaNet800 [8]	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2
RefineDet512 [37]	ResNet-101	36.4	57.5	39.5	16.6	39.9	51.4
CornerNet511 [13]	Hourglass-104	40.6	56.4	43.2	19.1	42.8	54.3
ExtremeNet511 [14]	Hourglass-104	40.2	55.5	43.2	20.4	43.2	53.1
CenterNet511 [12]	Hourglass-104	44.9	62.4	48.1	25.6	47.4	57.4
FoveaBox [10]	ResNeXt-101	42.1	61.9	45.2	24.9	46.8	55.6
FSAF [11]	ResNeXt-101	42.9	63.8	46.3	26.6	46.2	52.7
FCOS [9]	ResNet-101	41.0	60.7	44.1	24.0	44.1	51.0
FCOS [9]	ResNeXt-101	42.1	62.1	45.2	25.6	44.9	52.0
ours	ResNet-101	43.4	62.0	46.5	24.8	46.9	55.1
ours	ResNeXt-101	44.7	63.6	48.4	26.1	48.5	57.1

Visualization

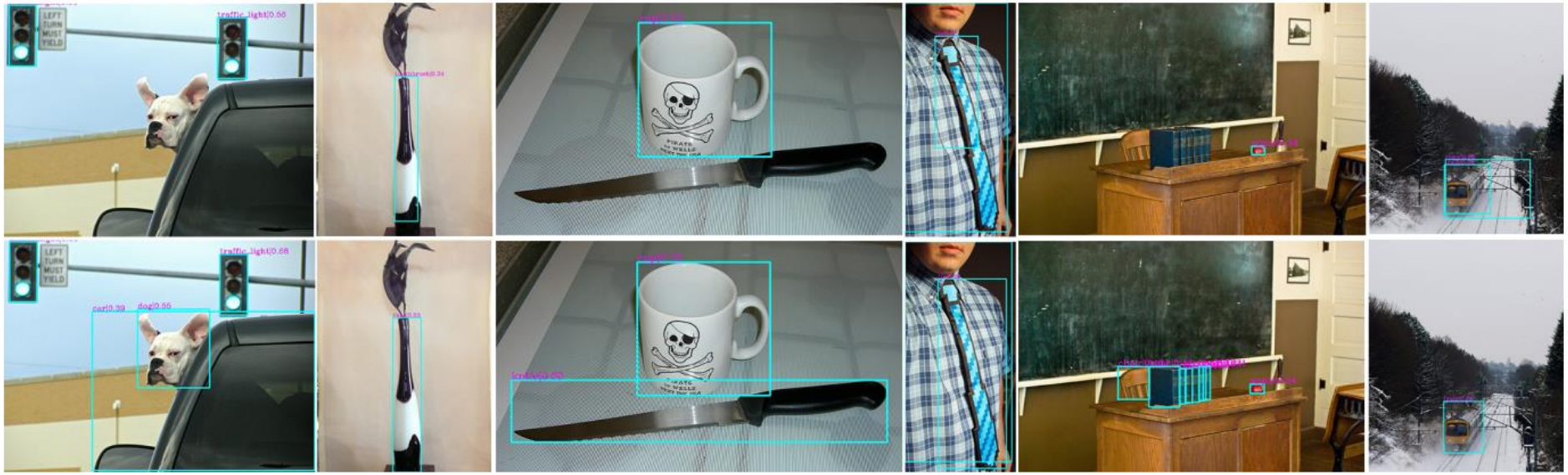


Fig. 6. Some comparison examples between FCOS (top) and our detector with BMFPN (bottom). Both are using ResNet-50 as backbone. Our BMFPN helps finding more challenging objects.

TABLE II

DETECTION RESULTS OF SOME CATEGORIES ON MS COCO VAL-2017 SPLIT. THE NUMBERS IN PARENTHESIS STANDS FOR THE RELATIVE AP IMPROVEMENT.

Method	airplane	snowboard	surfboard	fork	keyboard	toaster	scissors	toothbrush	refrigerator	tennis racket
FCOS	61.6	21.4	26.0	23.1	42.8	21.7	21.3	14.3	49.0	43.2
ours	69.5(**+7.9**)	32.3(**+10.9**)	34.7(**+8.7**)	39.7(**+6.6**)	50.8(**+8.0**)	36.4(**+14.7**)	32.0(**+10.7**)	22.5(**+8.2**)	55.1(**+6.1**)	49.3(**+6.1**)

04

Conclusion

Conclusion

1. A poor match between rectangular objects and feature maps with square receptive filed.
2. A sparse information flow among each feature in the pyramid.
3. Bidirectional Matrix Feature Pyramid Network (BMFPN) is proposed to address these issues.
4. An end-to-end anchor-free detector is designed and trained by integrating BMFPN into FCOS.
5. Extensive experiments demonstrate the effectiveness of the proposed architecture and the novel modules.

上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

Thanks

饮 水 思 源　爱 国 荣 校