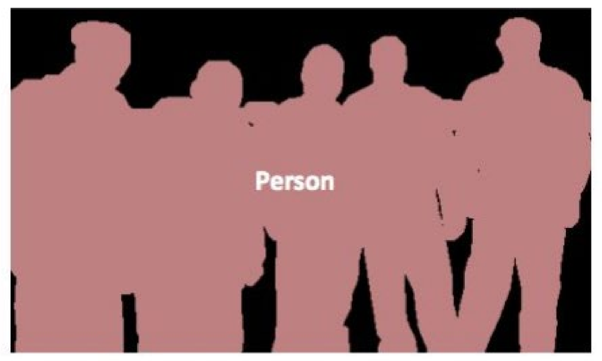
Accurate Instance Segmentation via Collaborative Learning

Tianyou Chen[1], Xiaoguang Hu[1], Jin Xiao[1], Guofeng Zhang[1], and Shaojie Wang[1]

[1]Beihang University

1. Introduction

Object Detection



Semantic Segmentation

Instance Segmentation

Problem:

Existing instance segmentation models:

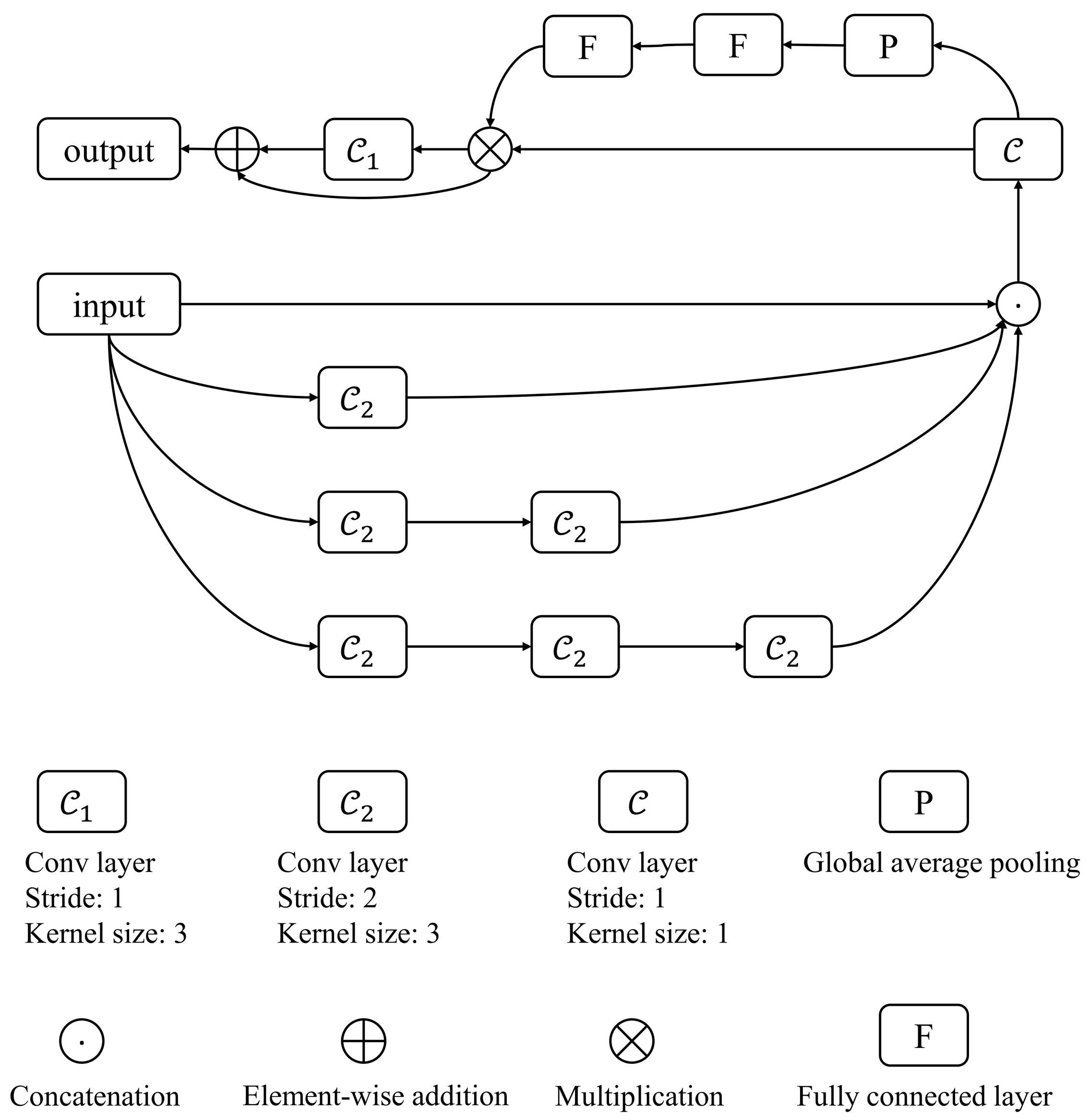
- Cannot solve the scale variation issue very well.
- Small effective receptive field size.
- Cannot fully leverage the foreground samples to train the regressor.

Contribution:

- We propose the MSFEM to exploit multi-scale spatial cues and enhance the single-level representation. Besides, the MSFEM can also enlarge the effective receptive field of the network, which is also helpful to improve the performance.
- We propose a collaborative learning framework where object detection and mask segmentation are integrated in a mutually beneficial manner.
- Extensive experimental results on the MS COCO dataset prove that the CoMask is competitive compared with state-of-the-art methods.

2. CoMask Model

a. Structure of the MSFEM.



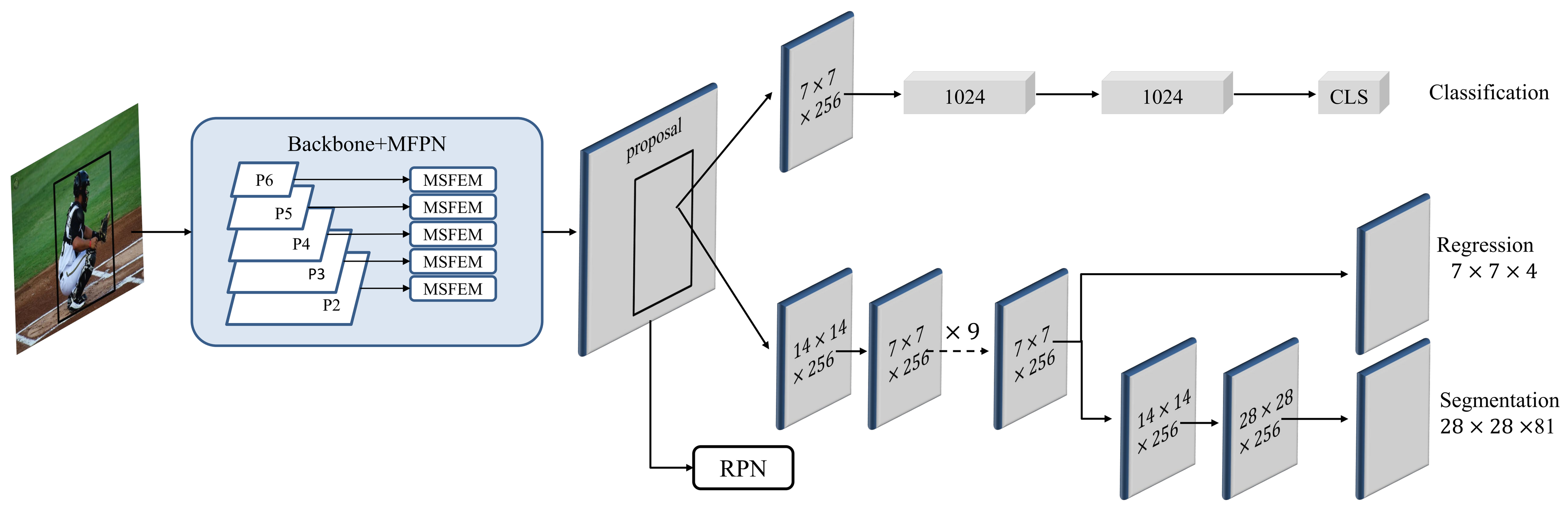
a. MSFEM

Each MSFEM contains four sub-branches and different sub-branches have different number of convolutional layers, which improve the performance with little computation overhead.

b. CoMask

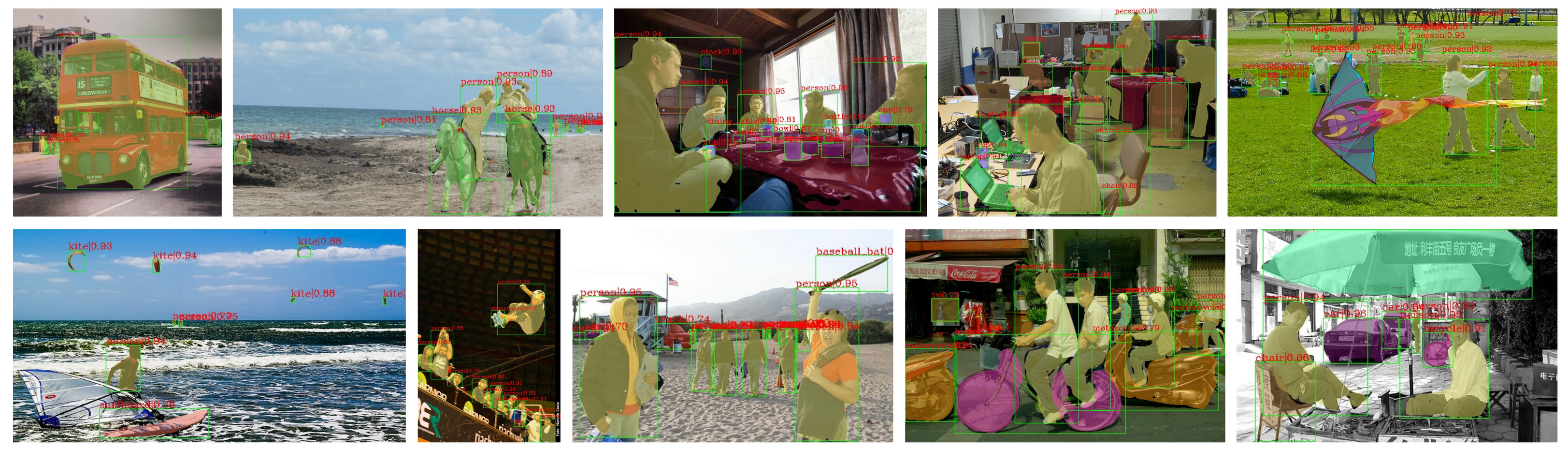
We innovatively integrate the object detection and mask segmentation in a mutually beneficial manner to avoid the interference of background regions on the final box regression

b. Overall architecture of the proposed CoMask. $\{P_2, \ldots, P_6\}$ are the output feature maps of the FPN.



3. Result

- Qualitative results of CoMask on COCO.



4. Comparison with SOTAs

- Best results are highlighted in **BOLDFACE**.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
PANet [10]	ResNet-50	36.6	58.0	39.3	16.3	38.1	**53.1**
CondInst [23]	ResNet-50	35.4	56.4	37.6	18.4	37.9	46.9
BlendMask [3]	ResNet-50	34.3	55.4	36.6	14.9	36.4	48.9
CoMask	ResNet-50	**37.7**	**59.0**	**40.9**	**21.0**	**40.9**	48.5
MS RCNN [2]	ResNet-101	38.3	58.8	41.5	17.8	40.4	**54.4**
Mask R-CNN [1]	ResNet-101	35.7	58.0	37.8	15.5	38.1	52.4
RetinaMask [25]	ResNet-101	34.7	55.4	36.9	14.3	36.7	50.5
ShapeMask [26]	ResNet-101	37.4	58.1	40.0	16.1	40.1	53.8
Cascaded Mask R-CNN [21]	ResNet-101	38.4	**60.2**	41.4	20.2	41.0	50.6
Mask SSD1024 [24]	ResNet-101	33.1	53.1	35.0	12.8	34.9	59.0
CoMask	ResNet-101	**38.6**	60.1	**41.9**	**21.2**	**41.9**	50.3

5. Ablation Studies

- $CoMask_4$ shows the best overall performance, which verifies the effectiveness of the proposed MSFEM. In particular, the improvement is more obvious when detecting large instances, which proves that MSFEM is effective in modeling larger context.
- As demonstrated in table 3, CoMask outperforms *w/o CL*, which validates the effectiveness of the collaborative learning framework.

Table 2. Ablation analysis for the proposed MSFEM. The inference speed of each variant is tested on a single NVIDIA Titan Xp Gpu. The best results are highlighted in **bold**.

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	FPS
$CoMask_1$	36.7	57.4	39.6	19.8	40.3	49.1	4.4
$CoMask_2$	37.2	**58.2**	**40.2**	20.1	**40.9**	49.9	4.2
$CoMask_4$	**37.3**	**58.2**	**40.2**	**20.2**	40.8	**50.3**	3.9

Table 3. Ablation analysis for the proposed collaborative learning framework. *w/o* CL indicates the variant without using collaborative learning strategy. The best results are highlighted in **bold**.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
CoMask	**37.3**	**58.2**	**40.2**	20.2	**40.8**	**50.3**
w/o CL	37.0	**58.2**	39.8	**20.3**	40.5	49.9