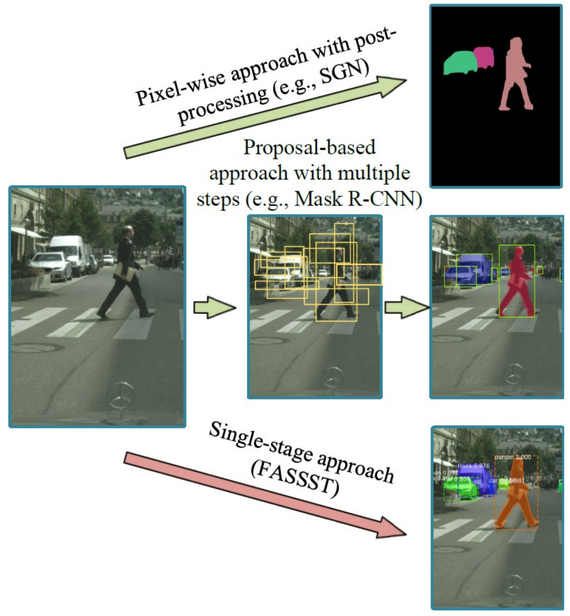
FASSST: Fast Attention Based Single-Stage Segmentation Net for Real-Time Instance Segmentation

Yuan Cheng[1,2], Rui Lin[2], Peining Zhen[1], Tianshu Hou[1], Chiu Wa Ng[2], Hai-Bao Chen[1], Hao Yu[3], Ngai Wong[2]

[1] Shanghai Jiao Tong University, [2] The University of Hong Kong, [3] Southern University of Science and Technology

Acknowledgement: Seed Funding, HKU-TCL Joint Research Centre for Artificial Intelligence, Hong Kong

1. Background

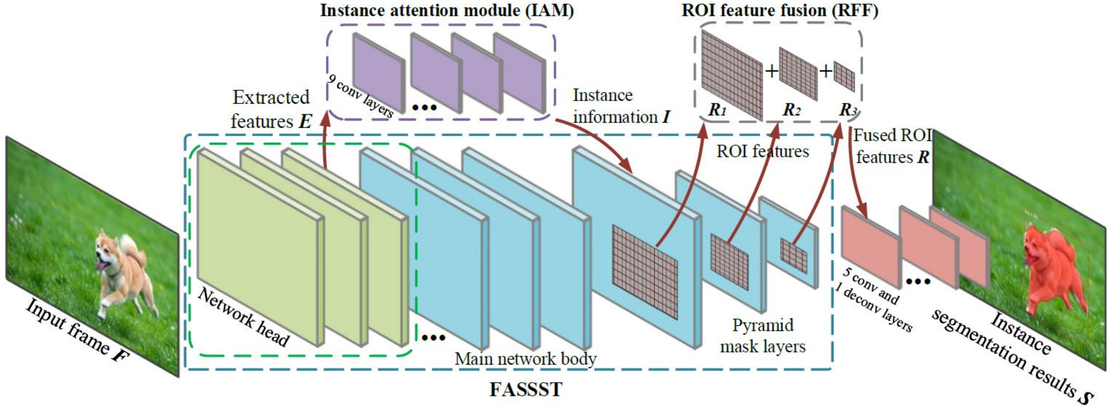


Modern researches on instance segmentation mainly fall into two categories:

- Pixel-wise approaches: learn an affinity relation between image pixels and segment image by segregating pixels of different instances and grouping pixels of the same instance.
- Proposal-based approaches: first propose object candidates by bounding boxes, then select interested ones of them, and perform masking at the end.

2. FASSST

We design FASSST (**F**ast **A**ttention-based **S**ingle-**S**tage **S**egmentation Ne**T**) for real-time instance segmentation.



FASSST quickly locates target instances and segments region of interest (ROI) by following steps:

- ***Step 1***, the raw feature tensor ***E*** is generated by the **network head**.
- ***Step 2***, ***E*** is parallelly delivered into the **main network body** and **instance attention module (IAM)**. The IAM regards the instance attention as a single-stage regression problem, which directly learns instance locality information ***I*** from raw features ***E***.
- ***Step 3***, instance locality information ***I*** is used to locate ROIs on several pyramid mask layers and obtain the fused ROI features ***R*** by an **ROI feature fusion module (RFF)**.
- ***Step 4***, the representation ***R*** is fed into the subsequent small-size convolutional layers to get the final instance segmentation results ***S***.

3. Experiments

Visual results on COCO

Accuracy comparison on COCO

Category	Approach	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Pixel-wise	SGN [22]	-	25.0	44.9	25.8	-	-	-
	SSAP [12]	ResNet-101-FPN	29.4	48.1	28.8	-	28.6	-
Proposal-based	FCIS [18]	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
	FCIS+++ [18]	ResNet-101-C5-dilated	33.6	54.5	**37.9**	-	-	-
	MNC [9]	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
	Mask R-CNN [13]	ResNet-101-FPN	**35.7**	**58.0**	37.8	**15.5**	**38.1**	52.4
Single-stage	ExtremeNet [32]	Hourglass-104	18.9	44.5	13.7	10.4	20.4	28.3
	YOLACT [2]	ResNet-101-FPN	31.2	50.6	32.8	12.1	33.3	47.1
	SOLO [28]	ResNet-101-FPN	**37.8**	**59.5**	**40.4**	**16.4**	**40.6**	**54.2**
	SipMask [17]	ResNet-101-FPN	32.8	53.4	34.3	9.3	35.6	**54.0**
	CenterMask [17]	ResNet-50-FPN	32.9	-	-	12.9	34.7	48.7
	PolarMask [30]	ResNet-101-FPN	30.4	51.9	31.0	13.4	32.4	42.8
Proposed	**FASSST**	MobileNet-54-V2	**34.2**	**56.4**	**38.1**	**14.9**	**36.7**	**53.8**

Main References

He K, et al., "Mask r-cnn", in Proceedings of the IEEE International Conference on Computer Vision. 2017: 2961-2969.

Wang X, et al., "Solo: Segmenting objects by locations", in European Conference on Computer Vision. 2020: 649-665.