1.Introduction

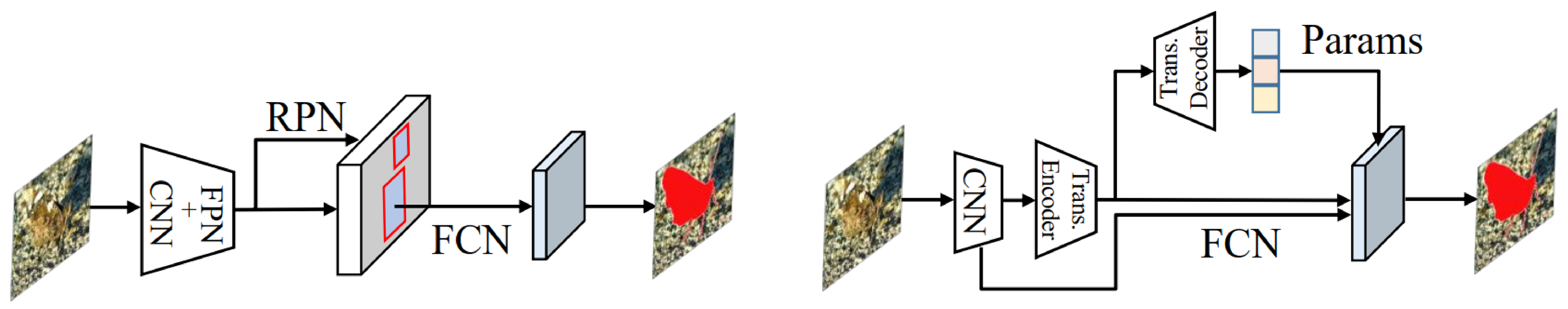
OSFormer: One-Stage Camouflaged Instance Segmentation with Transformers

Jialun Pei[1], Tianyang Cheng[1], Deng-Ping Fan[2], He Tang[1], Chuanbo Chen[1] and Luc Van Gool[2]

[1]Huazhong University of Science and Technology [2]ETH Zurich

Problems:

- COD only separates camouflages at region-level while ignoring **instance-level** identification.
- CIS needs to be performed in more complex scenarios with **high feature similarity** and results in **class-agnostic masks**.
- Camouflaged instances display **different camouflage strategies** in a scene, and they may combine to form **mutual camouflage**.
- The transformer-based model requires embracing **large-scale training data** and **longer training epochs**.

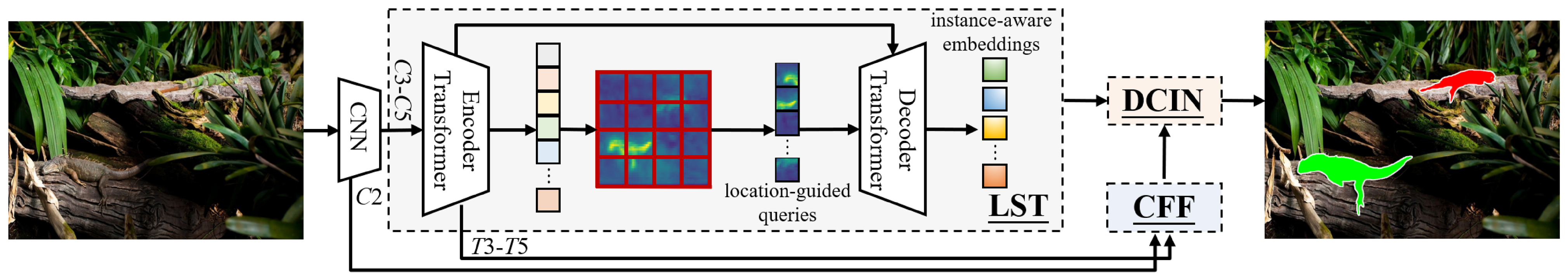


(a) Two-stage: Mask R-CNN (CNN+FPN) **(b) One-stage:** OSFormer (CNN+Transformer)

Contributions:

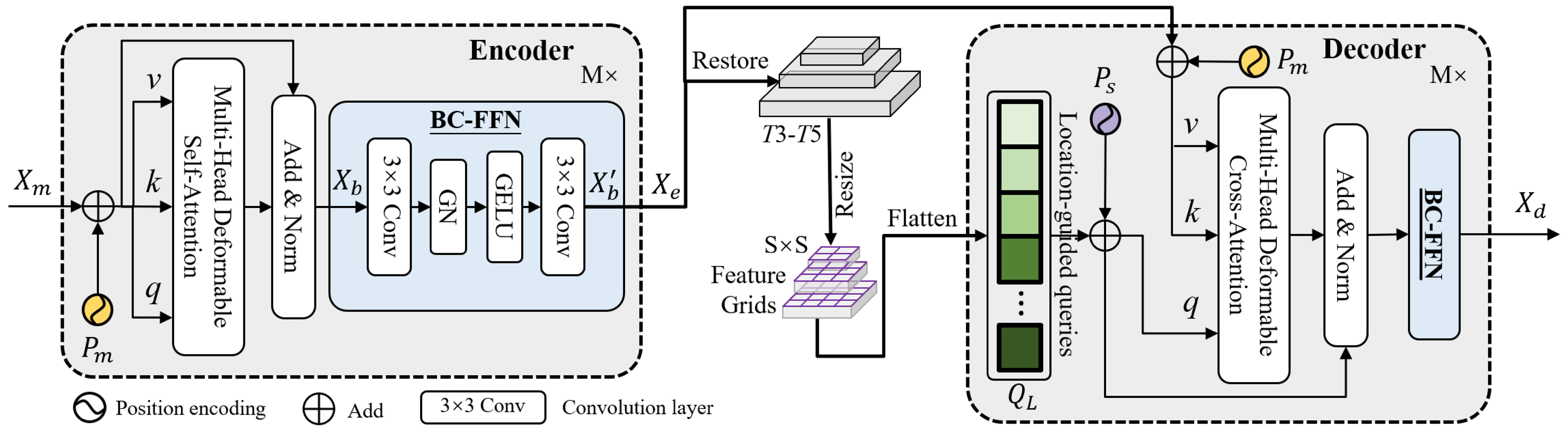
- Proposed **OSFormer**, the first **one-stage transformer-based framework** designed for the CIS task.
- Present a **Location-Sensing Transformer (LST)** to dynamically seize instance clues at different locations. LST contains an encoder with the **BC-FFN** and a decoder with the proposed **location-guided queries**.
- A novel **Coarse-to-Fine Fusion (CFF)** is proposed to get the high-resolution mask features. **Reverse edge attention (REA)** is embedded to highlight the edge information of instances
- OSFormer converges quickly with limited **3,000 training images**, outperforming 11 popular instance segmentation approaches by a large margin, **8.5% AP** improvement on the COD10K test set.

2.One-Stage Transformer for CIS (OSFormer)

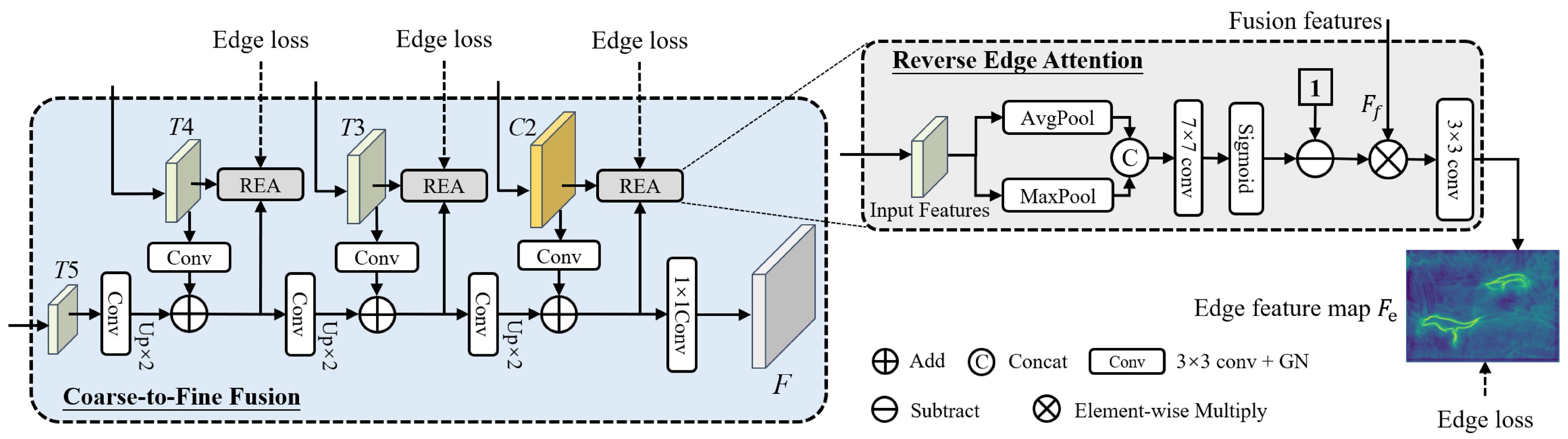


The proposed OSFormer comprises four essential components:

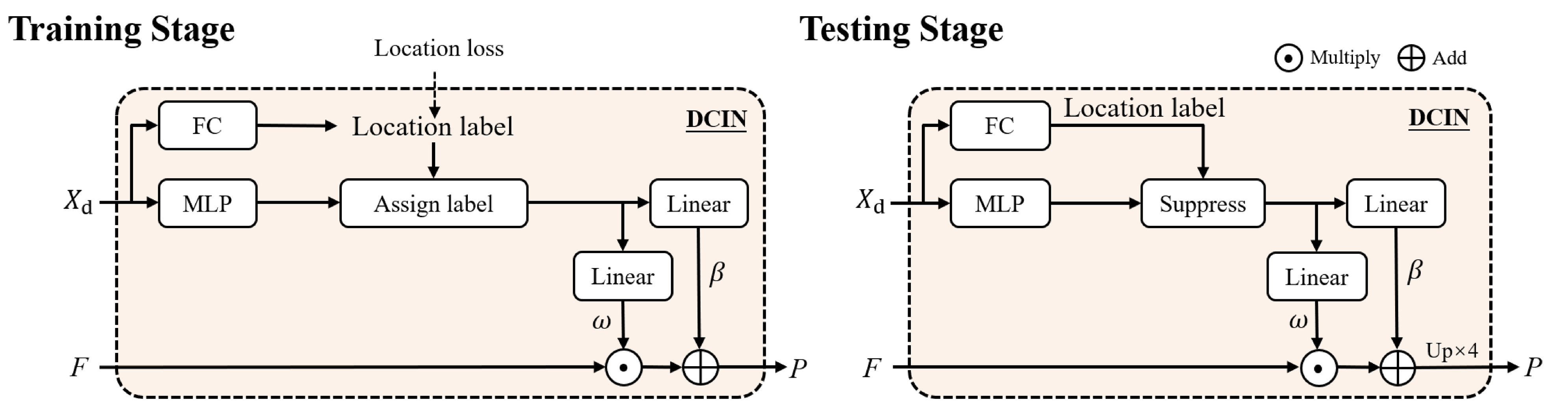
- A **CNN backbone** to extract object feature representation.
- A **location-sensing transformer (LST)** to produce the instance-aware embeddings.
- A **coarse-to-fine fusion (CFF)** to yield a high-resolution mask feature.
- A **dynamic camouflaged instance normalization (DCIN)** to predict the final masks.



Structure of our location-sensing transformer

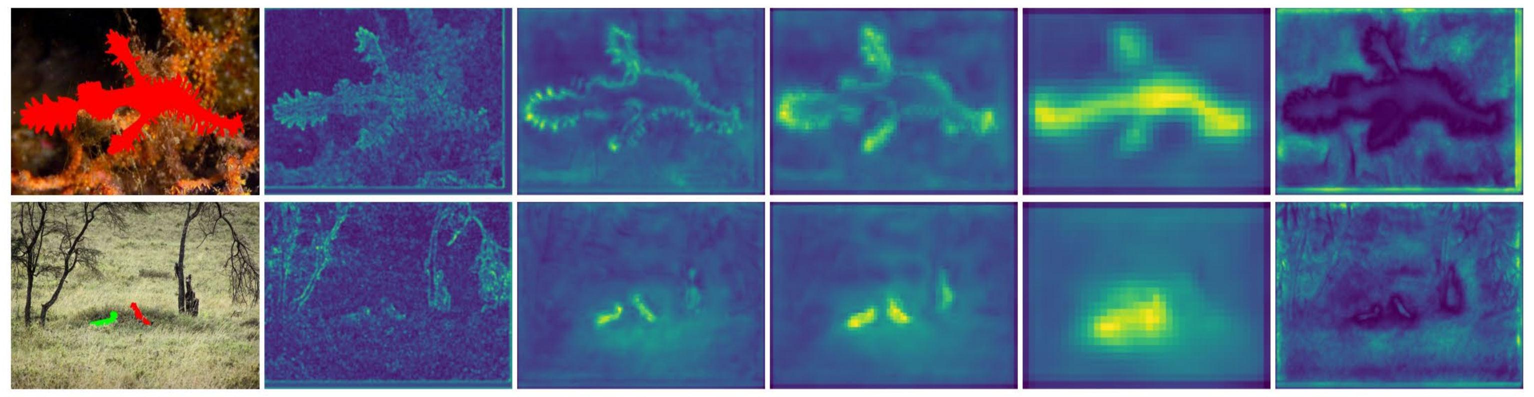


Structure of our coarse-to-fine fusion



Structure of our dynamic camouflaged instance normalization

3.Ablation Studies



(a) Image (b) $C2$ (c) $T3$ (d) $T4$ (e) $T5$ (f) F

Encoder	Decoder	AP	AP_{50}	AP_{75}	FPS
1	3	37.0	68.0	35.4	**21.8**
3	1	39.2	69.1	38.5	20.0
3	3	39.4	70.2	39.3	18.8
3	6	38.9	68.6	37.9	17.2
6	3	**41.0**	**71.1**	40.8	14.5
6	6	40.6	70.3	**41.2**	13.4
9	6	40.7	70.6	40.4	11.3

Queries	AP	AP_{50}	AP_{75}
Zero-Initialized [5]	34.7	64.1	33.1
Learnable Embeddings [65]	35.0	64.8	33.2
Location-Guided Queries (**Ours**)	**41.0** +6.0	**71.1** +6.3	**40.8** +7.6

Encoder	LGQ	BC-FFN	CFF	REA	AP	AP_{50}	AP_{75}
	✓	✓	✓	✓	33.7	63.4	32.0
✓		✓	✓	✓	34.7	64.1	33.1
✓	✓		✓	✓	37.2	67.3	35.8
✓	✓	✓			38.0	69.2	36.8
✓	✓	✓	✓		39.3	69.7	38.5
✓	✓	✓	✓	✓	**41.0**	**71.1**	**40.8**

4.Result

	Methods	Backbones	Params	FLOPs	COD10K-Test			NC4K-Test		
					AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Two-Stage	Mask R-CNN [23]	ResNet-50	43.9M	186.3G	25.0	55.5	20.4	27.7	58.6	22.7
	Mask R-CNN [23]	ResNet-101	62.9M	254.5G	28.7	60.1	25.7	36.1	68.9	33.5
	MS R-CNN [26]	ResNet-50	60.0M	198.5G	30.1	57.2	28.7	31.0	58.7	29.4
	MS R-CNN [26]	ResNet-101	79.0M	251.1G	33.3	61.0	32.9	35.7	63.4	34.7
	Cascade R-CNN [4]	ResNet-50	71.7M	334.1G	25.3	56.1	21.3	29.5	60.8	24.8
	Cascade R-CNN [4]	ResNet-101	90.7M	386.7G	29.5	61.0	25.9	34.6	66.3	31.5
	HTC [7]	ResNet-50	76.9M	331.7G	28.1	56.3	25.1	29.8	59.0	26.6
	HTC [7]	ResNet-101	95.9M	384.3G	30.9	61.0	28.7	34.2	64.5	31.6
	BlendMask [6]	ResNet-50	35.8M	233.8G	28.2	56.4	25.2	27.7	56.7	24.2
	BlendMask [6]	ResNet-101	54.7M	302.8G	31.2	60.0	28.9	31.4	61.2	28.8
	Mask Transfiner [29]	ResNet-50	44.3M	**185.1**G	28.7	56.3	26.4	29.4	56.7	27.2
	Mask Transfiner [29]	ResNet-101	63.3M	253.7G	31.2	60.7	29.8	34.0	63.1	32.6
One-Stage	YOLACT [3]	ResNet-50	-	-	24.3	53.3	19.7	32.1	65.3	27.9
	YOLACT [3]	ResNet-101	-	-	29.0	60.1	25.3	37.8	70.6	35.6
	CondInst [49]	ResNet-50	**34.1**M	200.1G	30.6	63.6	26.1	33.4	67.4	29.4
	CondInst [49]	ResNet-101	53.1M	269.1G	34.3	67.9	31.6	38.0	71.1	35.6
	QueryInst [19]	ResNet-50	-	-	28.5	60.1	23.1	33.0	66.7	29.4
	QueryInst [19]	ResNet-101	-	-	32.5	65.1	28.6	38.7	72.1	37.6
	SOTR [22]	ResNet-50	63.1M	476.7G	27.9	58.7	24.1	29.3	61.0	25.6
	SOTR [22]	ResNet-101	82.1M	549.6G	32.0	63.6	29.2	34.3	65.7	32.4
	SOLOv2 [57]	ResNet-50	46.2M	318.7G	32.5	63.2	29.9	34.4	65.9	31.9
	SOLOv2 [57]	ResNet-101	65.1M	394.6G	35.2	65.7	33.4	37.8	69.2	36.1
	OSFormer (Ours)	ResNet-50	46.6M	324.7G	41.0	71.1	40.8	42.5	72.5	42.3
	OSFormer (Ours)	ResNet-101	65.5M	398.2G	**42.0**	**71.3**	**42.8**	**44.4**	**73.7**	**45.1**

5.Visualization

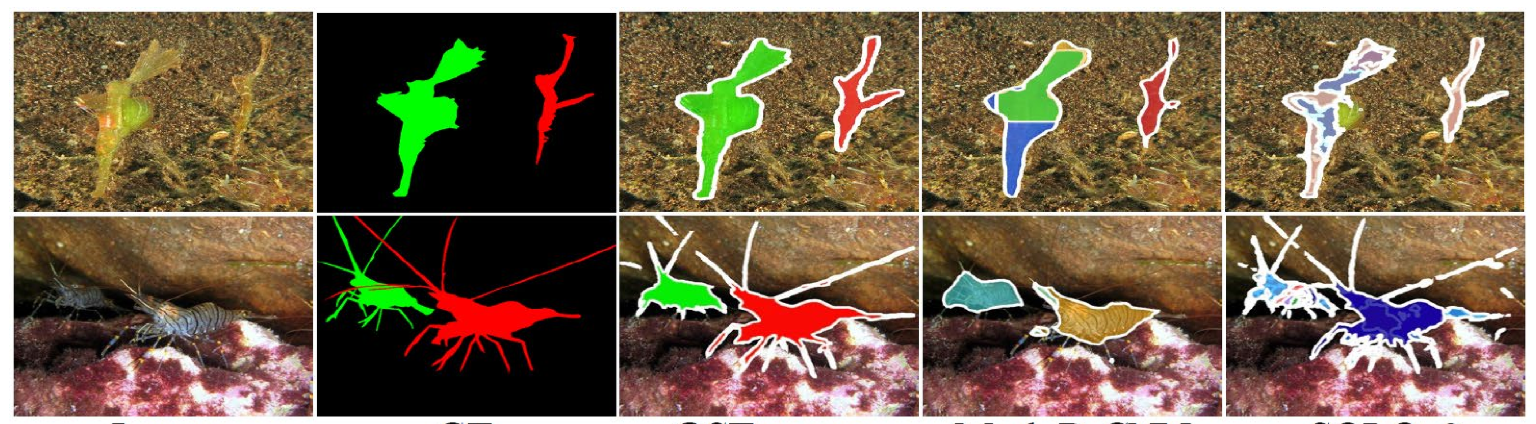


Image GT **OSFormer** Mask R-CNN SOLOv2

Paper, Code, and Result:
https://github.com/PJLallen/OSFormer