P2Seg: Pointly-supervised Segmentation via Mutual Distillation

Zipeng Wang[1*] Xuehui Yu[1*] Xumeng Han[1] Wenwen Yu[1] Zhixun Huang[2] Jianbin Jiao[1] and Zhenjun Han[1†]

[1]University of Chinese Academy of Sciences, Beijing, China [2]Xiaomi AI Lab, Beijing, China

[*] means：Equal Contribution；[†]means：Corresponding Author, hanzhj@ucas.ac.cn

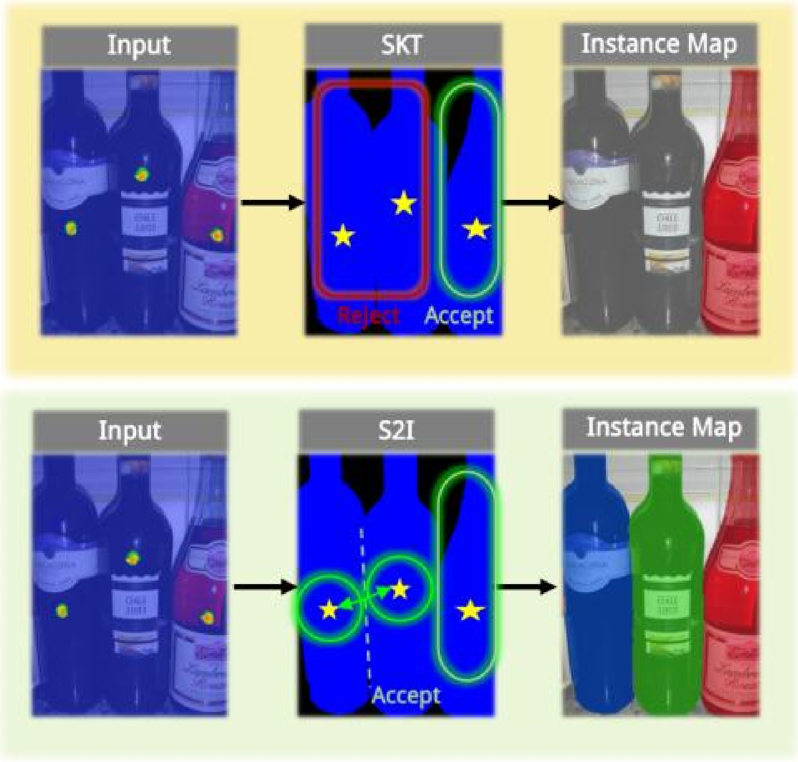
code：https://github.com/ucas-vg/P2Seg-Public/tree/main.

Introduction

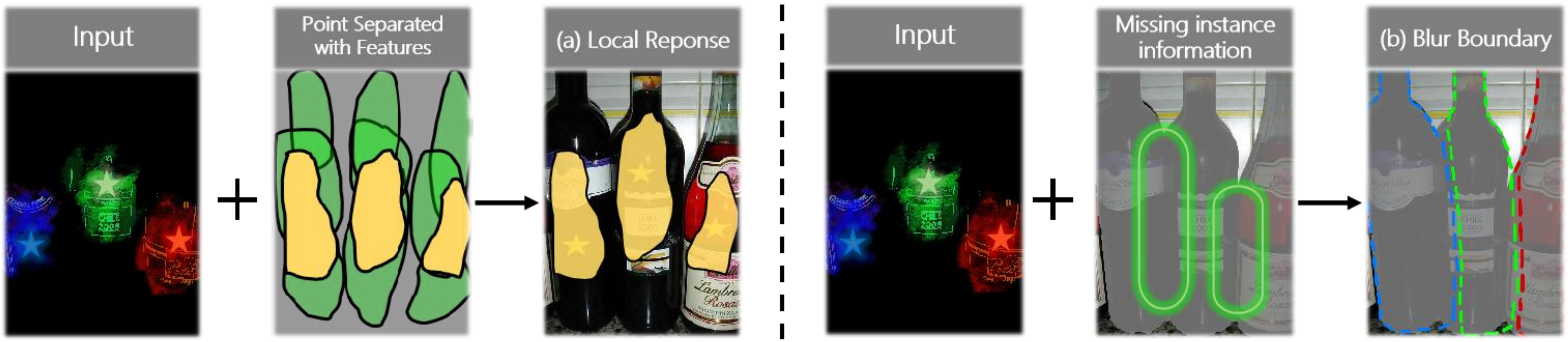
➢ Point-level Supervised Instance Segmentation (PSIS) aims to enhance the applicability and scalability of instance segmentation by utilizing low-cost yet instance informative annotations.

➢ Instance segmentation methods use point annotations that only approximate object positions, making it difficult to capture detailed features and accurate boundaries.

➢ Semantic segmentation excels at precise semantic region boundaries, it often struggles with instance-level discrimination within the same category.



Introduction



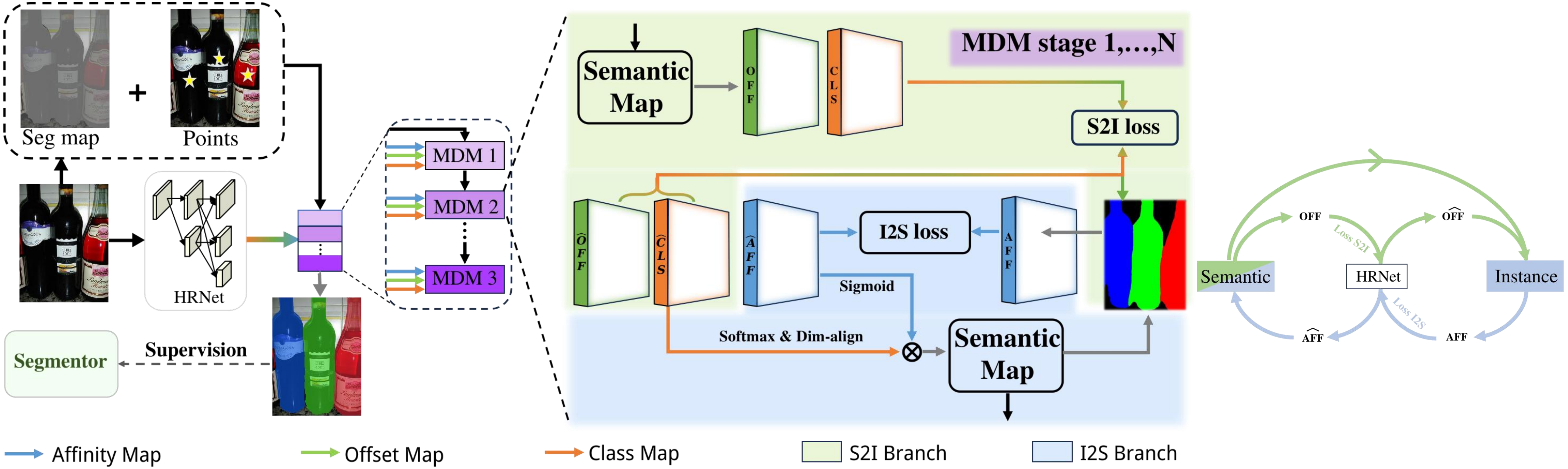
Existing PSIS methods usually rely on positional information to distinguish objects, but predicting precise boundaries remains challenging due to the lack of contour annotations, which suffers from the following problems:

- ➢ The local response caused by the separation of points from image features.
- ➢ The semantic segmentation estimation and instance differentiation are separated.

Table 8: Quantitative analysis for segmenting adjacent objects and addressing missing object issues.

Method	Missing Rate	Adjacent Rate
BESTIE	46.8	22.2
Ours	**42.9**	**52.6**

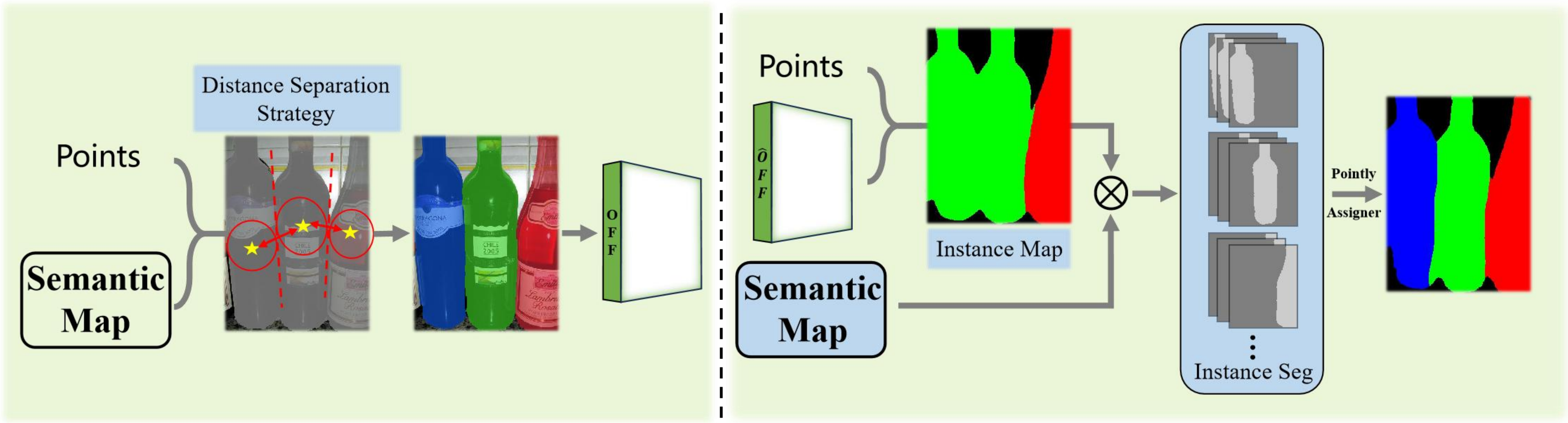
Methods(MDM)



Our MDM framework:(S2I branch is colored in green, I2S branch is colored in blue.)

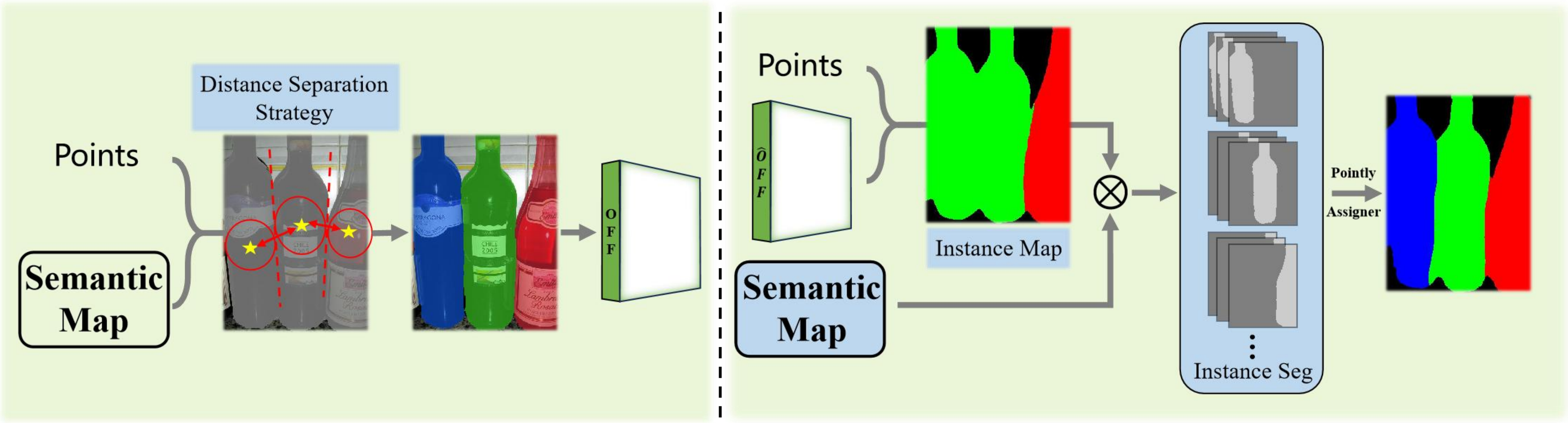
- In S2I branch, instance segmentation map is generated from the results of semantic segmentation using the offset map.
- In I2S branch, semantic segmentation results are influenced by instance segmentation map using affinity matrix.
- Training: $\mathcal{L} = \lambda_{I2S}\mathcal{L}_{I2S} + \lambda_{S2I}\mathcal{L}_{S2I}$

Methods(S2I)



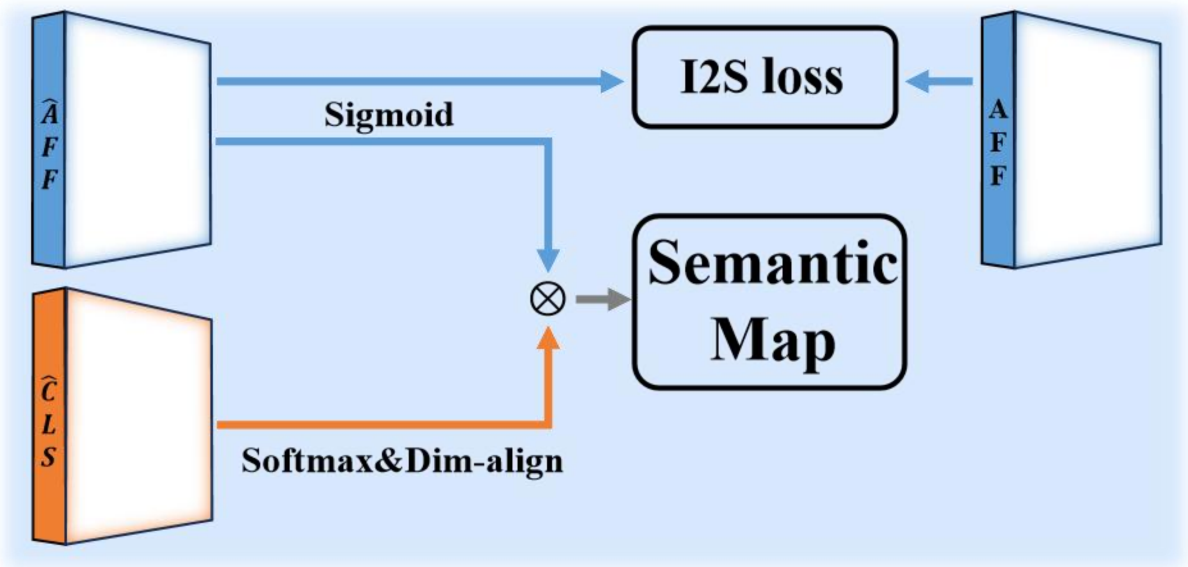
- Point annotations + Semantic Segmentation ⟶ initial Instance Segmentation labels;
- Distance separation strategy: Solve conflicts (e.g., multiple annotations in one region);
- Generates OFF maps & CLS maps.
- Class-agnostic Instance Segmentation map ⊗ Semantic Segmentation map (I2S branch) ⟶ New Instance Segmentation results.
- S2I Training: OFF maps(HR-Net) calculate the loss function with initial offset map.

Methods(S2I)



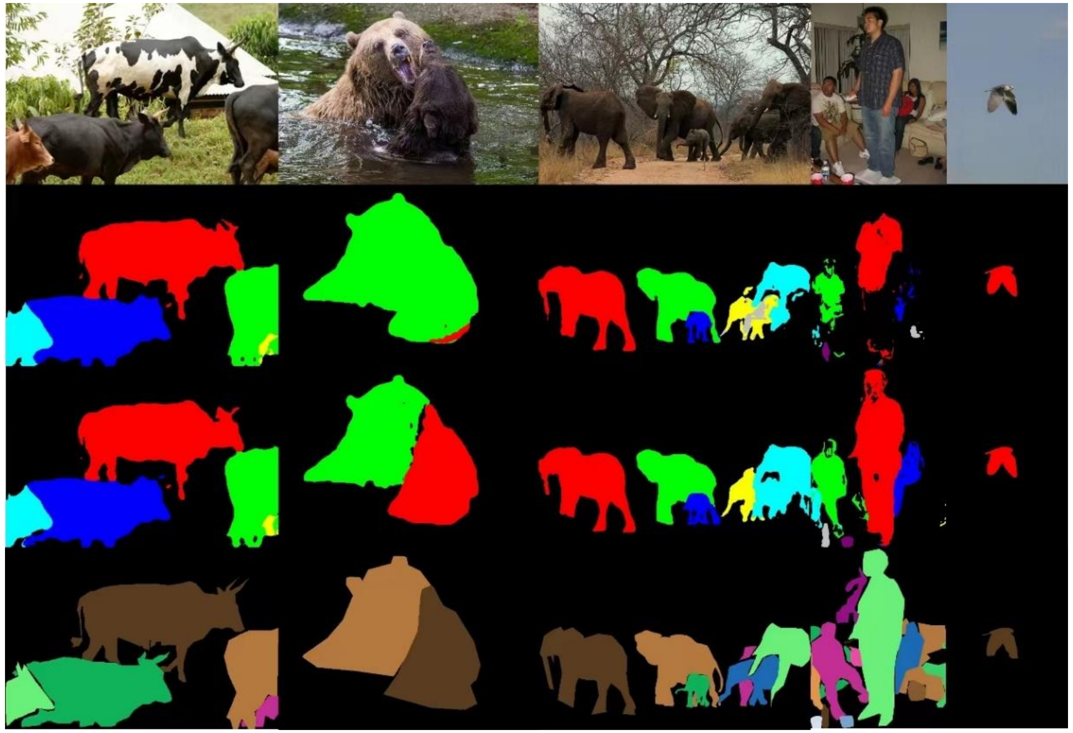
- Point annotations: $E = \{e_1, e_2, \cdots, e_K\}$;
- Distance separation strategy: $k^* = \arg\min_{k} \|e_k - p\|$;
- Generates OFF maps : $\mathcal{O} = \bigcup_{e_k \in E} \|e_k - m\|$.
- S2I Training: $\mathcal{L}_{S2I} = \mathcal{L}_{off} + \mathcal{L}_{seg}$; $\mathcal{L}_{off} = \frac{1}{|\mathcal{P}_{\text{pseudo}}|} \sum_{(i,j) \in \mathcal{P}_{\text{pseudo}}} \mathcal{W}(i,j) \cdot \text{smooth}_{L1} |\hat{\mathcal{O}}(i,j) - \mathcal{O}(i,j)|$; $\mathcal{L}_{seg} = \frac{1}{|\mathcal{P}_{seg}|} \sum_{(i,j) \in \mathcal{P}_{seg}} \text{CE}(p_{\hat{\mathcal{C}}(i,j)}, \mathcal{C}(i,j))$

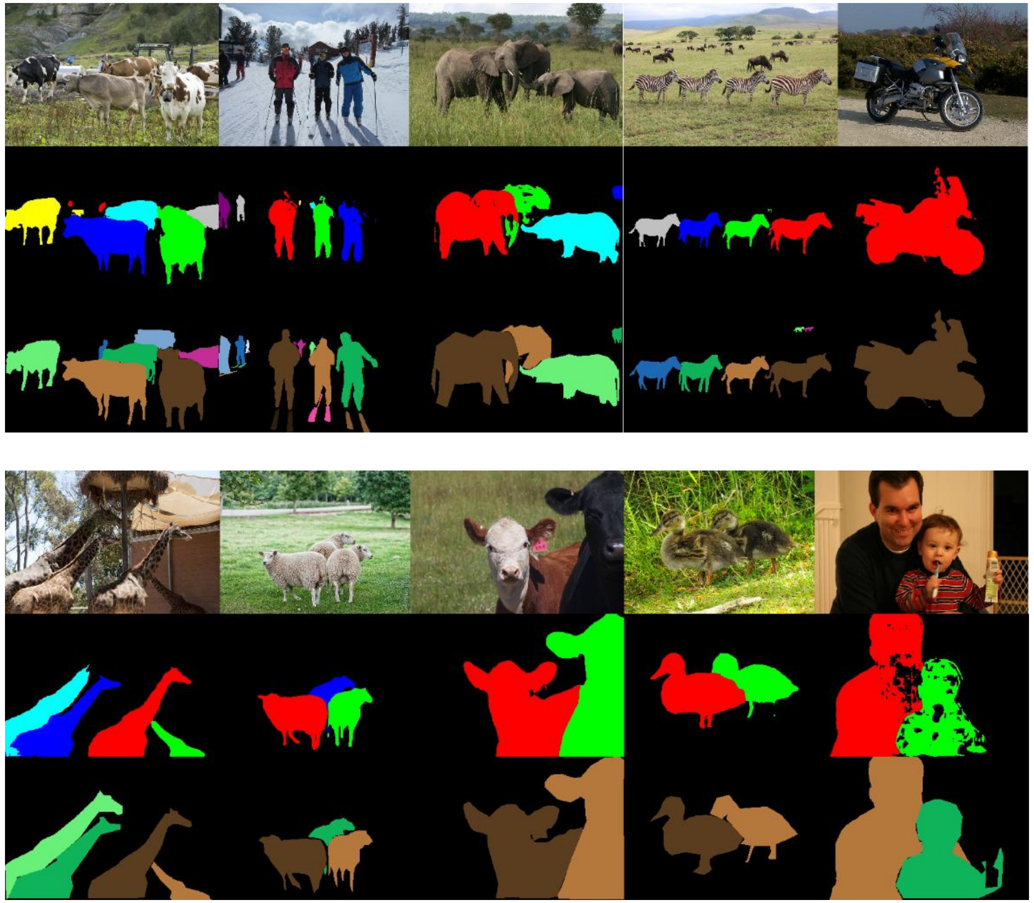
Methods(I2S)



- Obtains AFF matrix: two pixels (the same instance ☑: value=1, ☒: value=0);
- Generate the updated Semantic Segmentation map: $\mathcal{S} = \mathcal{A}^{\circ\beta} * \mathcal{C}$;
- I2S Training: Then AFF (Instance Segmentation map) is used to compute the $\mathcal{L}_{I2S}$ with the predicted $\widehat{AFF}$(HR-Net): $\mathcal{L}_{I2S} = \frac{1}{N^+}\sum_{(i,j)\in R^+}\left(2 - \sigma(\mathcal{A}^{ij}) - \sigma(\widehat{\mathcal{A}}^{ij})\right) + \frac{1}{N^-}\sum_{(k,l)\in R^-}\left(\sigma(\mathcal{A}^{kl}) + \sigma(\widehat{\mathcal{A}}^{kl})\right)$;
- Semantic segmentation map obtained through I2S serves as the Semantic information input for the next stage. Through the constraints of instances, I2S enriches the Semantic Segmentation results with instance information.

Visualization

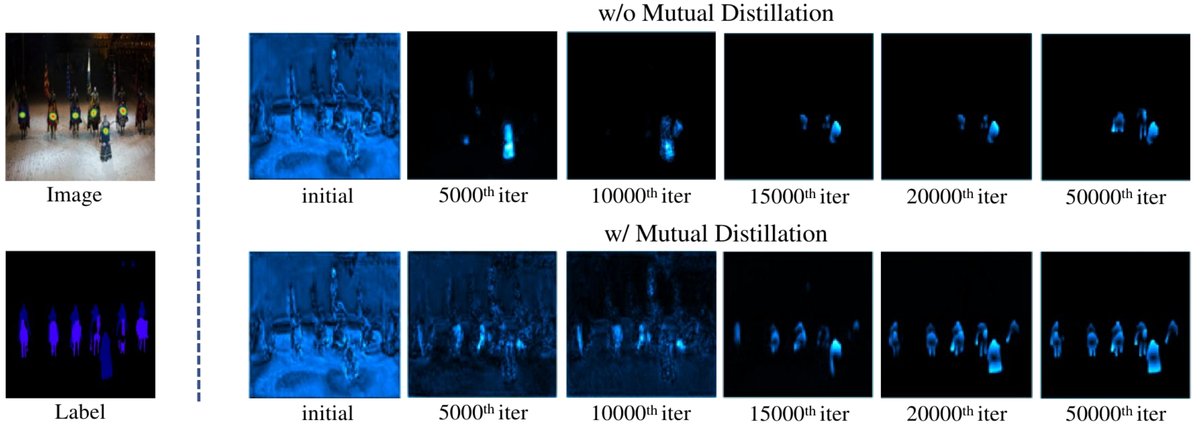




Visualization result comparison for our labels with ground truth on COCO.

Visualization

Visualization result comparison for our labels with ground truth on VOC.



Comparison of class-agnostic instance segmentation learning. **Left:** Original image and class-agnostic instance segmentation annotations. **Right:** Learning progression from initialization to 50,000 iterations. The top without mutual distillation, and the bottom with it.

Experiments

Table 1: Quantitative comparison of the state-of-the-art WSIS methods on VOC 2012 val-set. We denote the supervision sources as: $\mathcal{F}$ (full mask), $\mathcal{B}$ (box-level label), $\mathcal{I}$ (image-level label), $\mathcal{P}$ (point-level label), $\mathcal{S}$ prompting SAM with ViT-Base for object mask annotations and $\mathcal{C}$ (object count). The off-the-shelf proposal techniques are denoted as follows: $\mathcal{M}$ (segment proposal (Pont-Tuset et al., 2017)), $\mathcal{R}$ (region proposal (Uijlings et al., 2013)), and $\mathcal{S}_{\mathcal{I}}$ (salient instance segmentor (Fan et al., 2017)).

Method	*Sup.*	Backbone	Extra	mAP_{25}	mAP_{50}	mAP_{70}	mAP_{75}
Mask R-CNN (He et al., 2017a)	$\mathcal{F}$	ResNet-50	-	76.7	67.9	-	44.9
End-to-End weakly-supervised models.							
PRM (Zhou et al., 2018)	$\mathcal{I}$	ResNet-50	$\mathcal{M}$	44.3	26.8	-	9.0
IAM (Zhu et al., 2019)	$\mathcal{I}$	ResNet-50	$\mathcal{M}$	45.9	28.3	-	11.9
Label-PEnet (Ge et al., 2019)	$\mathcal{I}$	VGG-16	$\mathcal{R}$	49.2	30.2	-	12.9
CL (Hwang et al., 2021)	$\mathcal{I}$	ResNet-50	$\mathcal{M},\mathcal{R}$	56.6	38.1	-	12.3
BBTP (Hsu et al., 2019)	$\mathcal{B}$	ResNet-101	-	23.1	54.1	-	17.1
BBTP w/CRF	$\mathcal{B}$	ResNet-101	-	27.5	59.1	-	21.9
BoxInst (Tian et al., 2021)	$\mathcal{B}$	ResNet-101	-	-	60.1	-	34.6
OCIS (Cholakkal et al., 2019)	$\mathcal{C}$	ResNet-50	$\mathcal{M}$	48.5	30.2	-	14.4
Point2Mask Li et al. (2023)	$\mathcal{P}$	ResNet-101	-	-	48.4	-	22.8
Multi-Stage weakly-supervised models.							
WISE (Laradji et al., 2019)	$\mathcal{I}$	ResNet-50	$\mathcal{M}$	49.2	41.7	-	23.7
IRN (Ahn et al., 2019)	$\mathcal{I}$	ResNet-50	-	-	46.7	23.5	-
LIID (Liu et al., 2020)	$\mathcal{I}$	ResNet-50	$\mathcal{M},\mathcal{S}_{\mathcal{I}}$	-	48.4	-	24.9
Arun *et al.* (Arun et al., 2020)	$\mathcal{I}$	ResNet-101	$\mathcal{M}$	59.7	50.9	30.2	28.5
WISE-Net (Laradji et al., 2020)	$\mathcal{P}$	ResNet-50	$\mathcal{M}$	53.5	43.0	-	25.9
BESTIE† (Kim et al., 2022)	$\mathcal{P}$	ResNet-101	-	60.8	52.3	-	30.3
BESTIE† (Kim et al., 2022)	$\mathcal{P}$	HRNet-48	-	62.8	52.8	-	31.2
SAM (Kirillov et al., 2023)	$\mathcal{P}+\mathcal{S}$	ViT-S/22.1M	-	59.4	39.9	-	19.0
Ours	$\mathcal{P}$	ResNet-101	-	**63.1**	**53.9**	**37.7**	**32.0**
Ours	$\mathcal{P}$	HRNet-48	-	**66.0**	**55.6**	**40.2**	**34.4**

Compare with the state-of-the-art on VOC dataset

Experiments

Table 2: Quantitative comparison of the state-of-the-art WSIS methods on MS COCO 2017 val-set. We denote the supervision sources as: $\mathcal{F}$ (full mask), $\mathcal{B}$(box-level label), $\mathcal{I}$(image-level label), and $\mathcal{P}$ (point-level label). The off-the-shelf proposal techniques are denoted as follows: $\mathcal{M}$ (segment proposal (Pont-Tuset et al., 2017)).

Method	*Sup.*	Backbone	Extra	AP	AP_{50}	AP_{75}
Mask R-CNN (He et al., 2017a)	$\mathcal{F}$	ResNet-50	-	34.6	56.5	36.6
End-to-End weakly-supervised models.						
BBTP (Hsu et al., 2019)	$\mathcal{B}$	ResNet-101	-	21.1	45.5	17.2
BoxInst (Tian et al., 2021)	$\mathcal{B}$	ResNet-101	-	31.6	54.0	31.9
Point2Mask (Li et al., 2023)	$\mathcal{P}$	ResNet-101	-	12.8	26.3	11.2
Multi-Stage weakly-supervised models.						
IRN (Ahn et al., 2019)	$\mathcal{I}$	ResNet-50	-	6.1	11.7	5.5
WISE-Net (Laradji et al., 2020)	$\mathcal{P}$	ResNet-50	$\mathcal{M}$	7.8	18.2	8.8
BESTIE† (Kim et al., 2022)	$\mathcal{P}$	HRNet-48	-	14.2	28.4	22.5
Ours	$\mathcal{P}$	HRNet-48	-	17.6	33.6	**28.1**

Method	*Sup.*	Extra	AP	AP_{50}	AP_{75}
COCO test-dev.					
Mask R-CNN (He et al., 2017a)	$\mathcal{F}$	-	35.7	58.0	37.8
Fan *et al.* (Fan et al., 2018)	$\mathcal{I}$	-	13.7	25.5	13.5
LIID (Liu et al., 2020)	$\mathcal{I}$	$\mathcal{M}$,$\mathcal{S}_{\mathcal{I}}$	16.0	27.1	16.5
BESTIE† (Kim et al., 2022)	$\mathcal{P}$	-	14.2	28.6	12.7
Ours	$\mathcal{P}$	-	**17.4**	**33.3**	**16.4**

Compare with the state-of-the-art on COCO dataset

Experiments

Table 3: Ablation study for our S2I and I2S, compared with BESTIE.

S→I	I→S	mAP_{50}
BESTIE		52.8
S2I		53.2
S2I	I2S	55.7

Table 4: Ablation study for different Segmentor Backbones.

Method	Segmentor	mAP_{50}
BESTIE	Mask-RCNN	52.8
Ours		55.6
BESTIE	SOLOv2	51.9
Ours		54.1

Table 5: Ablation experiment to analyze the impact of hard pixel ratio.

Hard pixel ratio	mAP_{50}
0.1	52.0
0.2	50.8
0.4	51.5
0.8	50.9

Table 6: The comparison of BESTIE and our P2Seg for IoU with the ground truth.

Method	IoU> 50	IoU>70	IoU>90	overall IoU
Semantic Results First	5782	4939	2440	58.49
Points First	**9544**	**7417**	**2558**	**66.57**

Ablation study (I) :

1. Effect of S2I and I2S on VOC.
2. Other Segmentors for Instance Segmentation.
3. Hard pixel ratio:“hard pixel ratio” refers to the proportion of challenging samples used in loss computation.
4. Prediction Mask Quality Comparison

Experiments

Table 8: Quantitative analysis for segmenting adjacent objects and addressing missing object issues.

Method	Missing Rate	Adjacent Rate
BESTIE	46.8	22.2
Ours	**42.9**	**52.6**

Table 9: Analysis of the effect of WSSS result on our WSIS performance.

Semantic Segmentation		Instance Segmentation
WSSS method	mIoU	mAP_{50}
PMM	70	55.6
Ground Truth	-	59.7

Table 10: Ablation experiment to analyze the impact of β.

β	mAP_{50}
1	51.5
2	51.0
3	51.0
4	51.0

Table 11: The ablation experiment to analyze the efficiency. Comparison of our method with the BESTIE method in terms of GFLOPs and FPS.

Method	GFLOPS	FPS
BESTIE	64.7	86.9 ms/img
Ours	**66.1**	**94.5 ms/img**

Table 12: Ablation study for our S2I and I2S, compared with BESTIE on COCO.

S→I	I→S	mAP_{50}
BESTIE		14.2
S2I		17.4
S2I	I2S	17.6

Table 13: Ablation experiment to analyze the impact of point drift. We apply Gaussian random perturbation to the coordinates of the center point of each object.

Drift point(σ)	mAP_{50}
Center point	64.4
5	64.4
10	63.8
15	62.9

Ablation study (II) :

1. Quantitative analysis
2. Influence of WSSS method.
3. The parameter β
4. Analyze the efficiency
5. Effect of S2I and I2S on COCO.
6. Drift-point: the center points drift

Thank you!