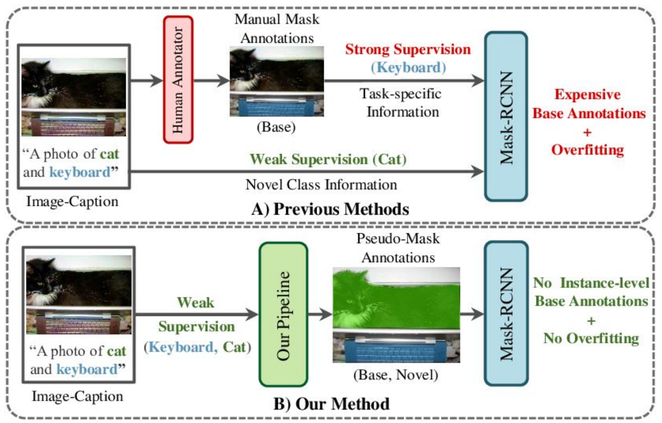
Mask-free OVIS: Open-Vocabulary Instance Segmentation without Manual Mask Annotations

Vibashan VS*, Ning Yu†, Chen Xing†, Can Qin‡, Mingfei Gao†,
Juan Carlos Niebles†, Vishal M. Patel*, Ran Xu†
Johns Hopkins University*, ‡Northeastern University, †Salesforce Research

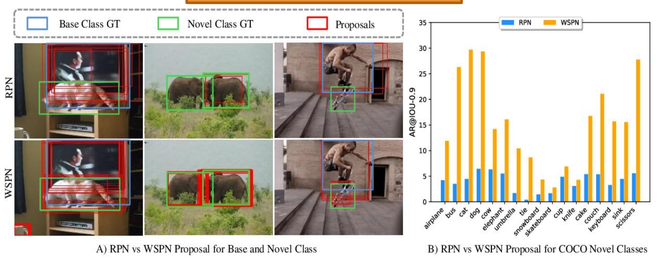
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Motivation



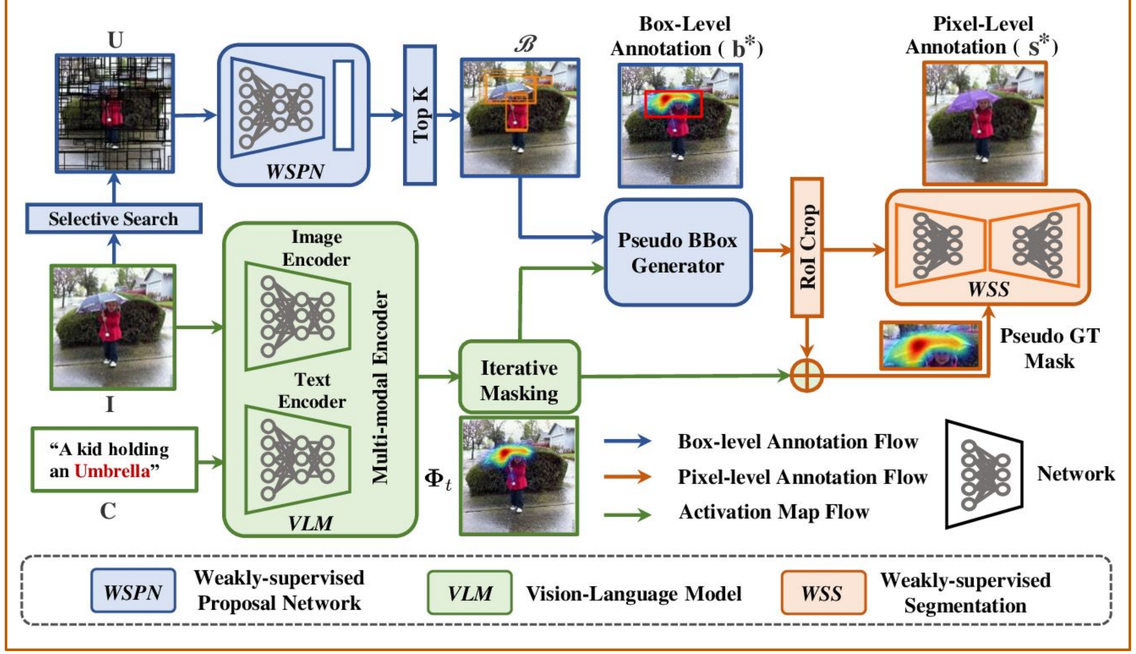
- Previous methods use fully-supervised learning for task-specific information, leading to overfitting and requiring expensive annotations.
- **Our method generates pseudo-annotations for both base and novel classes** under weak supervision, addressing the issues of costly annotation and overfitting.

Our Approach

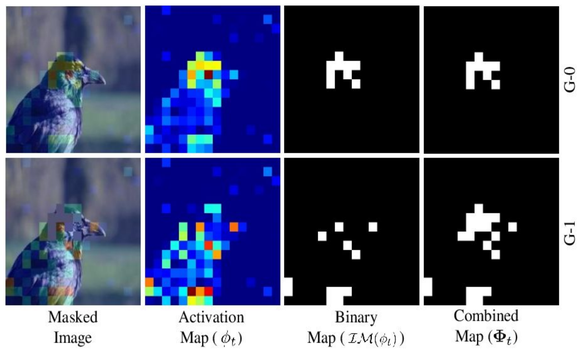


- RPN relies on fully-supervised learning with bounding box annotations, while WSPN uses weak supervision with image-labels.
- WSPN outperforms fully-supervised RPN by generating better quality proposals for novel object categories, mitigating the overfitting issue faced by RPN on base classes.

Network Architecture



Iterative Masking:



Pseudo-bbox Selection:

$$\mathbf{b}^* = \arg\max_{\mathbf{b}\in\mathcal{B}} \frac{\sum_{\mathbf{b}} \mathbf{\Phi}_t}{\sqrt{|\mathbf{b}|}}$$

Weakly-supervised Segmentation Loss:

$$\mathcal{L}_{wss} = \sum_{i=1}^{G} \mathcal{L}_{ce}(\mathbf{s}^*(f_i), \mathbf{\Theta}(f_i)) + \sum_{i=1}^{G} \mathcal{L}_{ce}(\mathbf{s}^*(b_i), \mathbf{\Theta}(b_i))$$

Region and Text embedding Similarity:

$$p(\mathbf{r}_i, \mathbf{c}_j) = \frac{\exp\left(h_{emb}(\mathbf{r}_i) \cdot \mathbf{c}_j\right)}{\exp(h_{emb}(\mathbf{r}_i) \cdot \mathbf{bg}) + \sum_k \exp\left(h_{emb}(\mathbf{r}_i) \cdot \mathbf{c}_k\right)}$$

Experiments and Results

Table 1. Object Detection (mAP) performances for MS-COCO under constrained and generalized setting. $\mathcal{C}_B$ and $\mathcal{C}_N$ are subset of $\mathcal{C}_\Omega$, where $\mathcal{C}_\Omega$ contains training vocabulary larger than COCO categories.

Method	Proposal Generator	Language Supervision	Base Annotation	Constrained Novel	Generalized Novel
WSDDN [5]	-	Image-labels in $\mathcal{C}_B \cup \mathcal{C}_N$	✗	-	19.7
Cap2Det [45]	-	Image-labels in $\mathcal{C}_B \cup \mathcal{C}_N$	✗	-	20.3
SB [2]	RPN $COCO_{base}$	-	✓	0.70	0.31
DELO [56]	RPN $COCO_{base}$	-	✓	7.60	3.41
PL [35]	RPN $COCO_{base}$	-	✓	10.0	4.12
OV-RCNN [46]	RPN $COCO_{base}$	Image-caption in $\mathcal{C}_B \cup \mathcal{C}_N$	✓	27.5	22.8
CLIP-RPN [14]	RPN $COCO_{base}$	CLIP image-text pair $\mathcal{C}_\Omega$	✓	-	26.3
ViLD [14]	RPN $COCO_{base}$	CLIP image-text pair $\mathcal{C}_\Omega$	✓	-	27.6
Detic [54]	RPN $COCO_{base}$	Image-caption in $\mathcal{C}_B \cup \mathcal{C}_N$	✓	-	27.8
RegionCLIP [52]	RPN $LVIS_{base}$	Conceptual caption $\mathcal{C}_\Omega$	✓	30.8	26.8
PB-OVD [11]	RCNN $COCO_{base}$	Image-caption in $\mathcal{C}_B \cup \mathcal{C}_N$	✓	32.3	30.7
XPM [17]	RPN $COCO_{base}$	Image-caption in $\mathcal{C}_B \cup \mathcal{C}_N$	✓	29.9	27.0
Mask-free OVIS (Ours)	WSPN $COCO_{base}$	Image-labels in $\mathcal{C}_B \cup \mathcal{C}_N$	✗	31.5	27.4
Mask-free OVIS (Ours)	WSPN $COCO_{base}$	Image-labels in $\mathcal{C}_B \cup \mathcal{C}_N$	✓	**35.9**	**31.5**

Table 2. Instance Segmentation (mAP) performances for MS-COCO and Open Images under constrained and generalized setting.

Method	Proposal Generator (MS-COCO/OpenImages)	Base Annotation	MS-COCO Constrained Novel	MS-COCO Generalized Novel	Open Images Constrained Novel	Open Images Generalized Novel
OVR+OMP [4]	-	✓	14.1	8.3	24.9	16.8
SB [2]	-	✓	20.8	16.0	24.8	17.3
BA-RPN [51]	-	✓	20.1	15.4	25.3	16.9
Soft-Teacher [43]	RPN $COCO_{base}$/RPN $OpenImg_{base}$	✓	14.8	9.6	25.9	17.6
Unbiased-Teacher [30]	RPN $COCO_{base}$/RPN $OpenImg_{base}$	✓	15.1	9.8	22.2	14.5
OV-RCNN [46]	RPN $COCO_{base}$/RPN $OpenImg_{base}$	✓	20.9	17.1	23.8	17.5
XPM [17]	RPN $COCO_{base}$/RPN $OpenImg_{base}$	✓	24.0	21.6	31.6	22.7
Mask-free OVIS (Ours)	WSPN $COCO_{base}$/WSPN $COCO_{base}$	✗	**27.4**	**25.0**	**35.9**	**25.8**