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

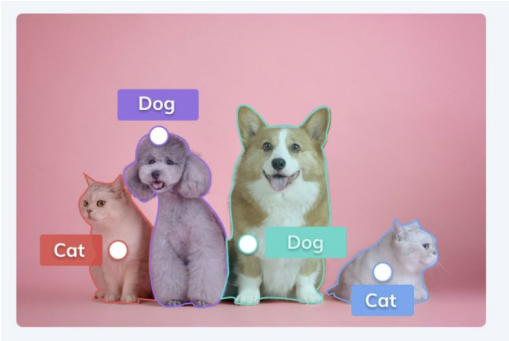
Vibashan VS*, Ning Yu†, Chen Xing†, Can Qin‡, Mingfei Gao†,
Juan Carlos Niebles†, Vishal M. Patel*, Ran Xu†
*Johns Hopkins University, ‡Northeastern University, †Salesforce Research

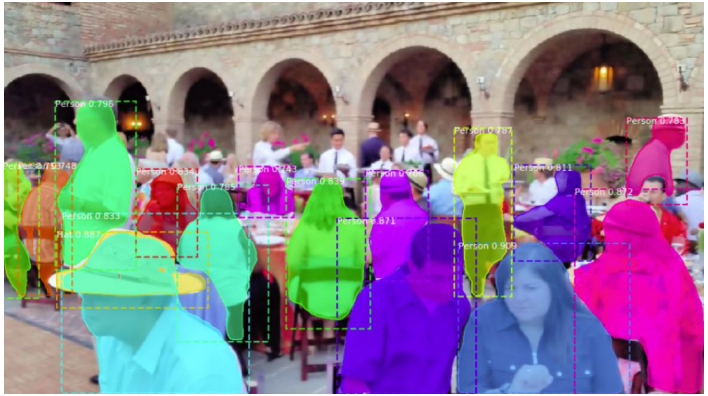
- Problem / objective
  - Instance segmentation
- Contribution / Key idea
  - Open-Vocabulary instance segmentation without manual mask annotations

VS, Vibashan, et al. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 23539-23549.

Instance segmentation

- 이미지 내에서 개별 객체를 픽셀 단위로 분할하고 동일한 카테고리 내에서도 각 객체를 구별하는 task
- 학습시 필요한 GT : 각 instance 의 바운딩박스 및 마스크 레이블
- 문제점 : 비싼 labeling cost 로 novel category 로의 확장 어려움.





VS, Vibashan, et al. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 23539-23549.

Open-Vocabulary (OV) instance segmentation

- Base category : Strong supervision, Novel category : Weak supervision (이미지-캡션 쌍을 통해 novel category 정보 학습)
- 문제점 :
  - Base category 에 오버피팅됨
  - Base category 레이블 여전히 요구

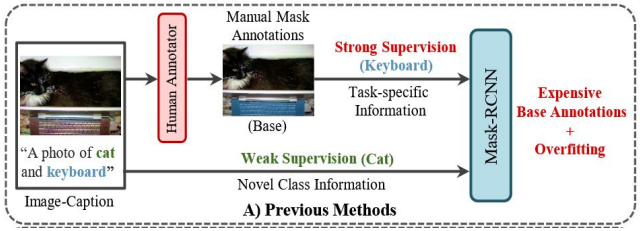
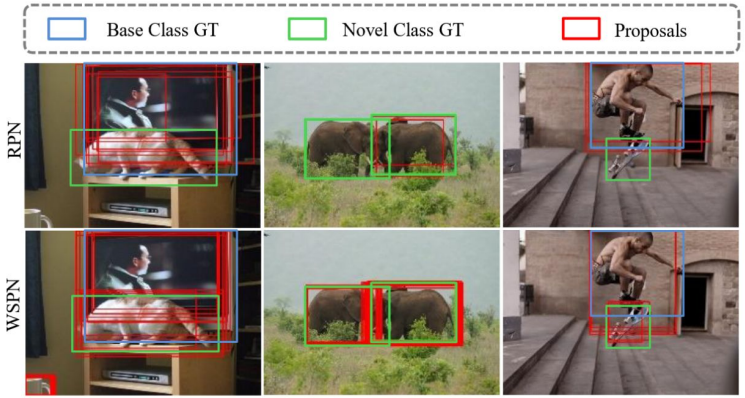
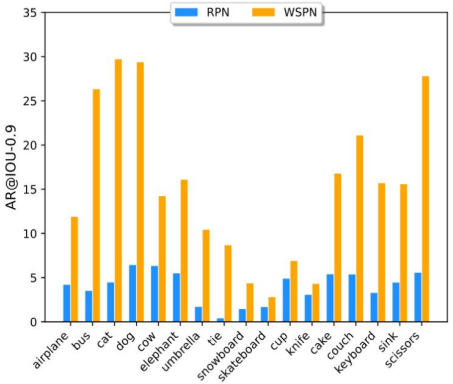


Figure 1. **A) Previous Methods:** Learn task-specific information (detection/segmentation) in a fully-supervised manner and novel category information with weak supervision. During training, this difference in strong and weak supervision signals leads to overfitting and requires expensive base annotations.



A) RPN vs WSPN Proposal for Base and Novel Class



B) RPN vs WSPN Proposal for COCO Novel Classes

Figure 2. RPN is supervised using bounding box annotations from COCO base and WSPN is supervised using image-labels from COCO base. **A)** WSPN produces better quality proposals for novel object categories compared to fully-supervised RPN. **B)** WSPN consistently produces better recall for all COCO novel categories than RPN.

VS, Vibashan, et al. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 23539-23549.

Mask-free OVIS

- 사람의 수고 없이 각 instance 의 레이블 생성하는 파이프라인 제안

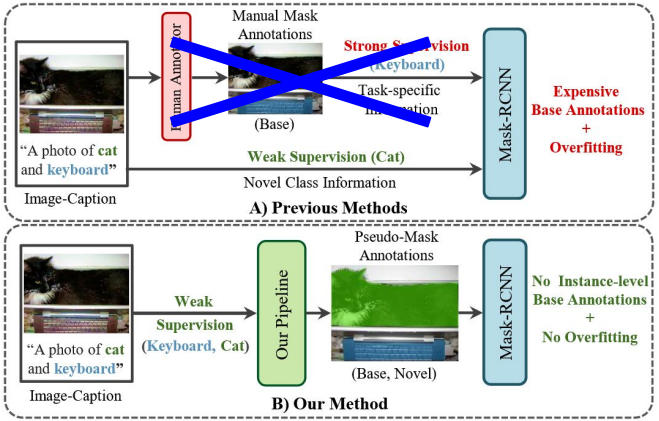


Figure 1. **A) Previous Methods:** Learn task-specific information (detection/segmentation) in a fully-supervised manner and novel category information with weak supervision. During training, this difference in strong and weak supervision signals leads to overfitting and requires expensive base annotations. **B) Our method:** Given image-caption pairs, we generate pseudo-annotations for both base and novel categories under weak supervision, solving the problems of labour-expensive annotation and overfitting.

VS, Vibashan, et al. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 23539-23549.

Overview

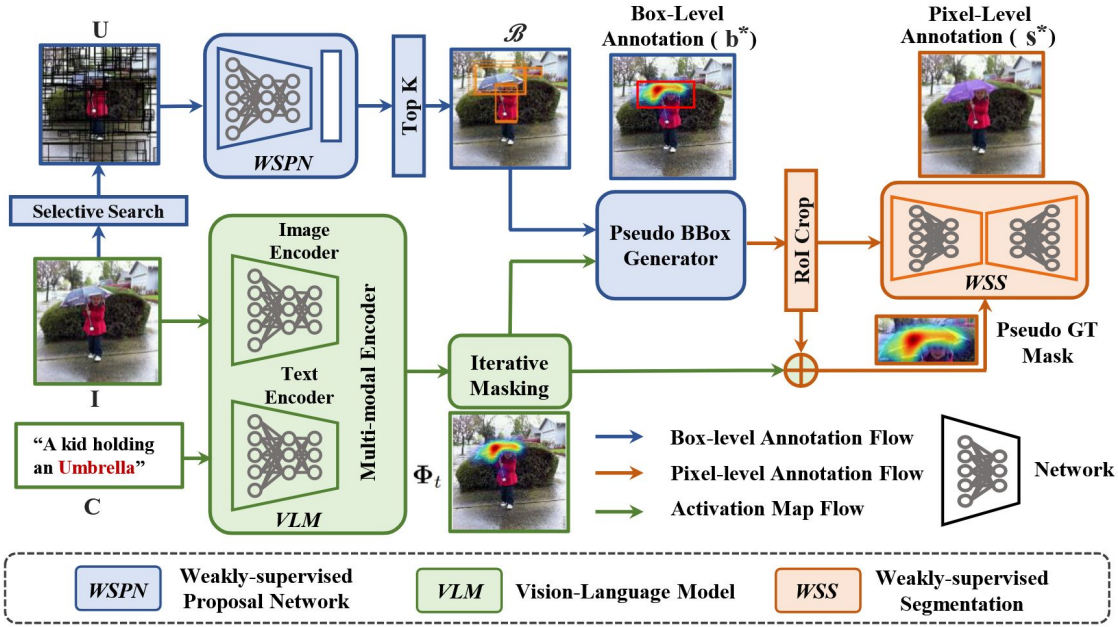


Figure 3. Illustrative overview of our pseudo-mask generation pipeline. Given an image-caption pair and pre-trained VLM, we generate an activation map for the object of interest ("umbrella") and enhance it using iterative masking strategy. We generate box-level annotations using an activation map as a guidance function to select the best WSPN proposals covering the object. We crop the image corresponding to the generated pseduo bounding box and perform weakly-supervised segmentation to obtain pixel-level annotations.

VS, Vibashan, et al. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 23539-23549.

GradCAM activation map generation

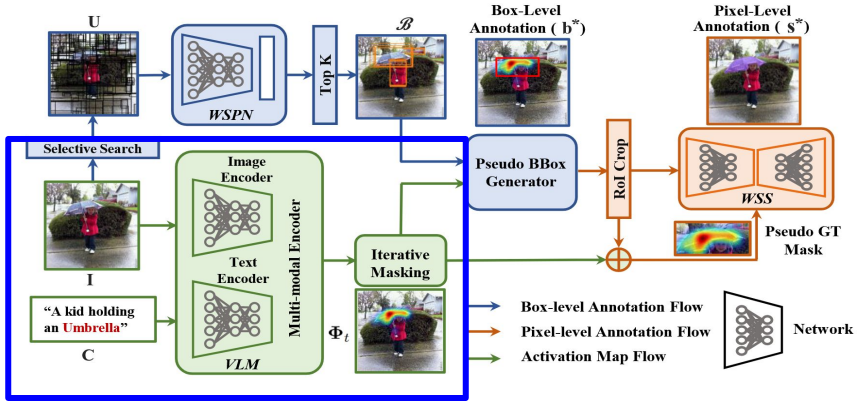


Image $\mathbf{I}$
Caption $\mathbf{C} = \{c_1, c_2, ..., c_{N_c}\}$
Region representations $\mathbf{R} \in \mathbb{R}^{N_R \times d}$
Text representations $\mathbf{T} \in \mathbb{R}^{N_C \times d}$
Hidden representations from previous $(m-1)$-th cross-attention layer $\mathbf{h}_t^{m-1}$
Image-caption similarity output from the multi-modal encoder's final layer $\mathbf{S}$

1. VLM의 m-th cross-attention layer 에서 캡션 내 객체 c_t 에 대한 attention score $\mathbf{X}_t^m$

$$\mathbf{X}_t^m = \text{Softmax}(\frac{\mathbf{h}_t^{m-1}\mathbf{R}^T}{\sqrt{d}}), \quad (1)$$

$$\mathbf{h}_t^n = \mathbf{X}_t^m \cdot \mathbf{R}. \quad (2)$$

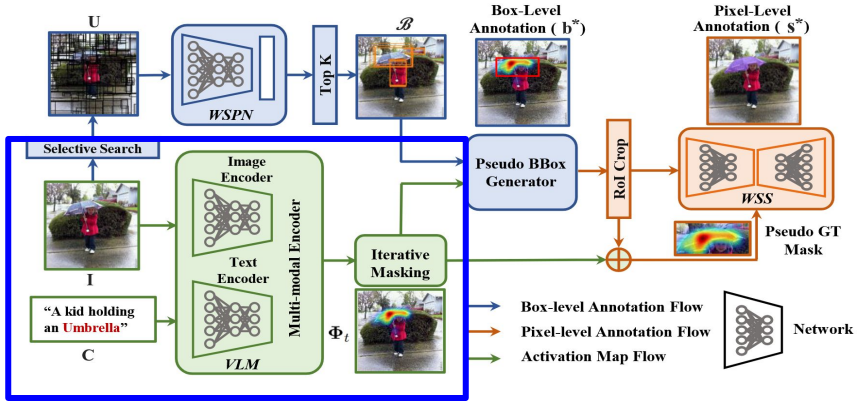
2. 객체 c_t 에 대한 activation map ϕ_t

$$\phi_t = \mathbf{X}_t^m \cdot \max\left(\frac{\partial S}{\partial \mathbf{X}_t^m}, 0\right). \quad (3)$$

전유진

VS, Vibashan, et al. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 23539-23549.

GradCAM activation map generation



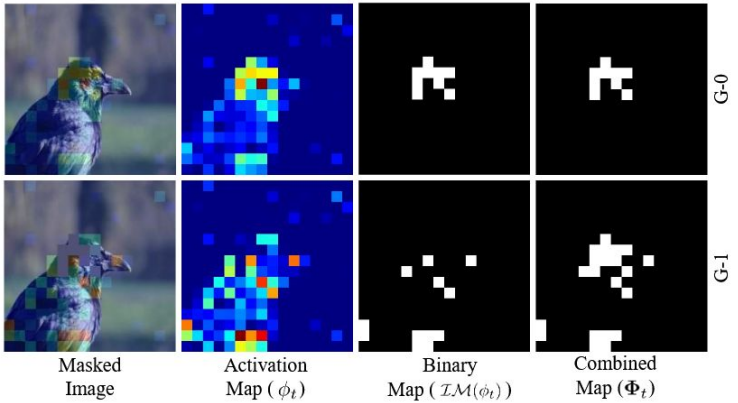


Figure 4. Comparison between activation map $\mathbf{\Phi}_t$ generated at G 0 and 1. For G=0, the most discriminative parts of an object (bird's head) gets activated (bird's head). After masking, for G=1 we can observe that the activation map has shifted to less discriminative part (bird's body). Thus, by combining activation from both steps, we obtain a better activation map trying to cover entire object $\mathbf{\Phi}_t$.

3. Iterative masking 을 통해 최종 activation map $\mathbf{\Phi}_t$ 생성

$$\mathbf{\Phi}_t = \bigcup_{i=1}^{G} \mathcal{IM}(\phi_t^i) \qquad (4)$$

VS, Vibashan, et al. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 23539-23549.

Pseudo bounding box generation

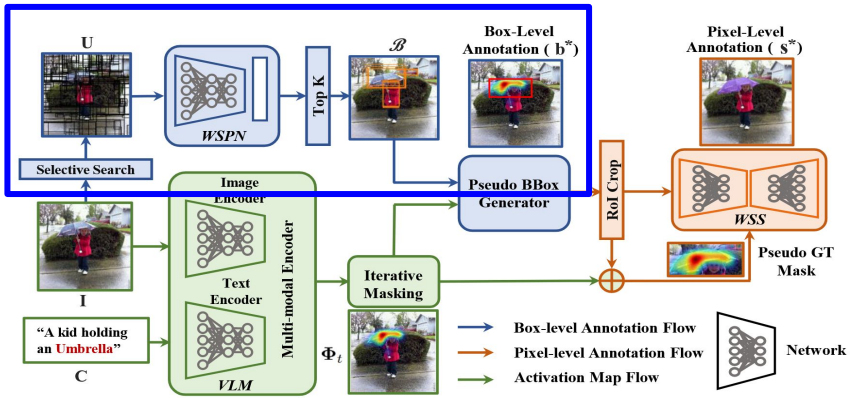


Image $\mathbf{I}$
Label $\mathbf{Y} = \{y_1, y_2, ..., y_C\}$
$y_c \Leftrightarrow 0$ or 1 (presence or absence of class c in $\mathbf{I}$)
Pseudo-regression targets for low-scoring proposals
$\hat{\mathbf{T}} = \{\hat{t}(u_1), \hat{t}(u_2), ..., \hat{t}(u_N)\}$

1. Selective Search 를 통해 bounding box proposals 생성

$$\text{Image } \mathbf{I} \rightarrow \text{bbox proposals } \mathbf{U} = \{u_1, u_2, ..., u_N\}$$

2. Weakly Supervised Proposal Network 학습

$\mathbf{W}^{cls}, \mathbf{W}^{det} \in \mathbb{R}^{C \times N}$: output of classification, detection branch for ROI pooled features

$\mathbf{W_C} = \sigma(\mathbf{W}^{cls}) \odot \sigma(\mathbf{W}^{det})$

$p_c = \sum_{i=1}^{N} w_{i,c}$: 각 이미지의 클래스별 classification score

[1] Bilen, Hakan, and Andrea Vedaldi. "Weakly supervised deep detection networks." *Proceedings of the IEEE conference on computer vision and pattern recogni*

VS, Vibashan, et al. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 23539-23549.

Pseudo bounding box generation

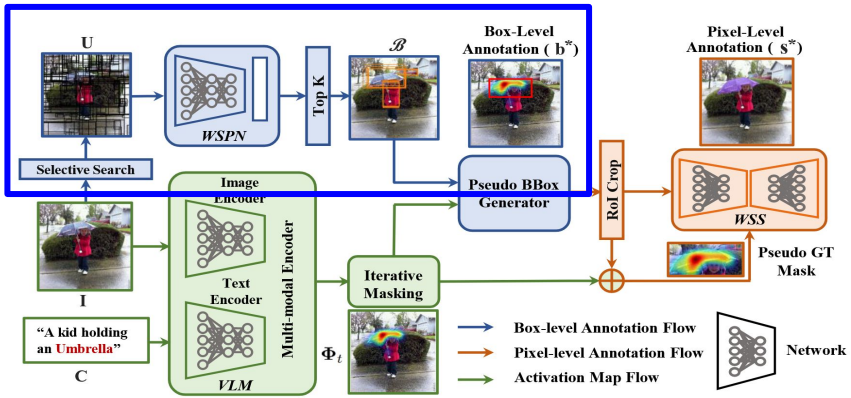


Image $\mathbf{I}$
Label $\mathbf{Y} = \{y_1, y_2, ..., y_C\}$
$y_c \Leftrightarrow 0$ or 1 (presence or absence of class c in $\mathbf{I}$)
Pseudo-regression targets for low-scoring proposals
$\hat{\mathbf{T}} = \{\hat{t}(u_1), \hat{t}(u_2), ..., \hat{t}(u_N)\}$

$$\mathcal{L}_{wspn} = -\sum_{c=1}^{C} y_c \log p_c + (1 - y_c) \log(1 - p_c) + \frac{1}{N} \sum_{u=1}^{N} \mathcal{L}_{smoothL1}(\hat{t}(u_i), u_i). \quad (5)$$

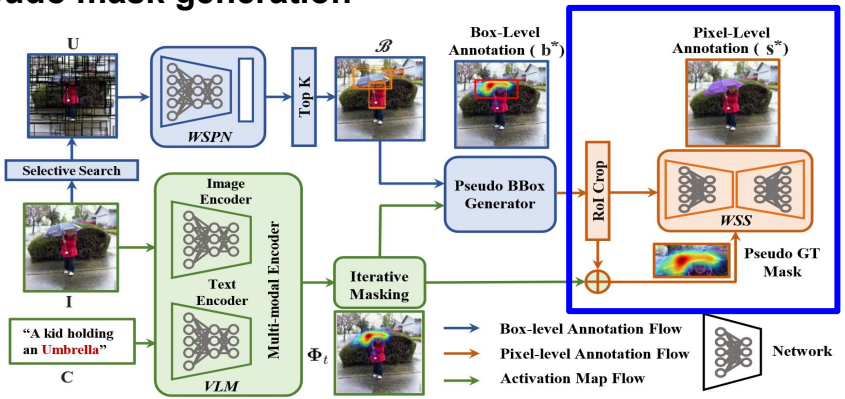
3. 학습된 모델 사용하여 생성한 object proposals 중 top-K 개 선택 $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_K\}$

4. Activation map 과 가장 많이 겹치는 proposal 을 box-level annotation 으로 지정 $\mathbf{b}^* = \arg\max_{\mathbf{b} \in \mathcal{B}} \frac{\sum_{\mathbf{b}} \mathbf{\Phi}_t}{\sqrt{|\mathbf{b}|}}, \quad (6)$

전유진

VS, Vibashan, et al. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 23539-23549.

Pseudo mask generation



1. Weakly Supervised Segmentation Network 학습

- 모델 : 3 layer CNN 구조
- 인풋 : cropped patch
- 레이블 : pseudo ground-truth

$$F_z = \{f_i\}_{i=1}^{Z} \rightarrow 1 : \text{most activated part of } \mathbf{\Phi}_t \text{ inside } \mathbf{b}^*$$
$$B_z = \{b_i\}_{i=1}^{Z} \rightarrow 0 : \text{least activated part of } \mathbf{\Phi}_t \text{ inside } \mathbf{b}^*$$

- 손실 : $$\mathcal{L}_{wss} = \sum_{i=1}^{G} \mathcal{L}_{ce}(\mathbf{s}^*(f_i), \mathbf{\Theta}(f_i)) + \sum_{i=1}^{G} \mathcal{L}_{ce}(\mathbf{s}^*(b_i), \mathbf{\Theta}(b_i)), \quad (7)$$

2. 학습된 모델의 예측 결과를 pixel-level annotation 으로 간주 $\mathbf{S}^*$

VS, Vibashan, et al. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 23539-23549.

Experiments

1. 데이터셋

- MS-COCO : 48 base categories + 17 novel categories

- Open Images : 200 base categories + 100 novel categories

2. 모델

- Vision Language 모델 : ALBEF

- Instance Segmentation 모델 : Mask R-CNN

[1] Li, Junnan, et al. "Align before fuse: Vision and language representation learning with momentum distillation." *Advances in neural information processing systems* 34 (2021): 9694-9705.
[2] He, Kaiming, et al. "Mask r-cnn." *Proceedings of the IEEE international conference on computer vision*. 2017.

VS, Vibashan, et al. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 23539-23549.

Experiments

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 [48]	-	Image-labels in $\mathcal{C}_B \cup \mathcal{C}_N$	✗	-	20.3
SB [2]	RPN $COCO_{base}$	-	✓	0.70	0.31
DELO [60]	RPN $COCO_{base}$	-	✓	7.60	3.41
PL [36]	RPN $COCO_{base}$	-	✓	10.0	4.12
OV-RCNN [49]	RPN $COCO_{base}$	Image-caption in $\mathcal{C}_B \cup \mathcal{C}_N$	✓	27.5	22.8
CLIP-RPN [15]	RPN $COCO_{base}$	CLIP image-text pair $\mathcal{C}_\Omega$	✓	-	26.3
ViLD [15]	RPN $COCO_{base}$	CLIP image-text pair $\mathcal{C}_\Omega$	✓	-	27.6
Detic [58]	RPN $COCO_{base}$	Image-caption in $\mathcal{C}_B \cup \mathcal{C}_N$	✓	-	27.8
RegionCLIP [56]	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 [18]	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**

- Constrained setting : test 이미지에 novel class 만 존재

- Generalized setting : test 이미지에 base class, novel class 둘다 존재

전유진

VS, Vibashan, et al. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 23539-23549.

Experiments

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		Open Images	
			Constrained Novel	Generalized Novel	Constrained Novel	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 [55]	-	✓	20.1	15.4	25.3	16.9
Soft-Teacher [46]	RPN $COCO_{base}$/RPN $OpenImg_{base}$	✓	14.8	9.6	25.9	17.6
Unbiased-Teacher [31]	RPN $COCO_{base}$/RPN $OpenImg_{base}$	✓	15.1	9.8	22.2	14.5
OV-RCNN [49]	RPN $COCO_{base}$/RPN $OpenImg_{base}$	✓	20.9	17.1	23.8	17.5
XPM [18]	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**

VS, Vibashan, et al. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 23539-23549.

Qualitative results

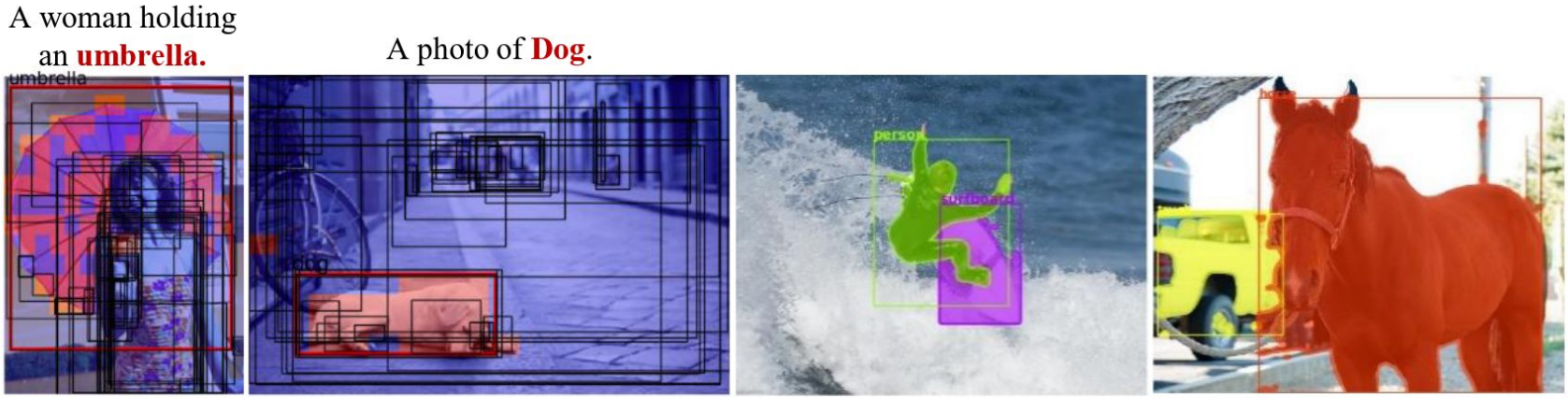


Figure 5. A) Pseduo bounding box selection guided by GradCAM activation. B) Visualization of pseudo-mask generated for Open Images.

VS, Vibashan, et al. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 23539-23549.

Qualitative results

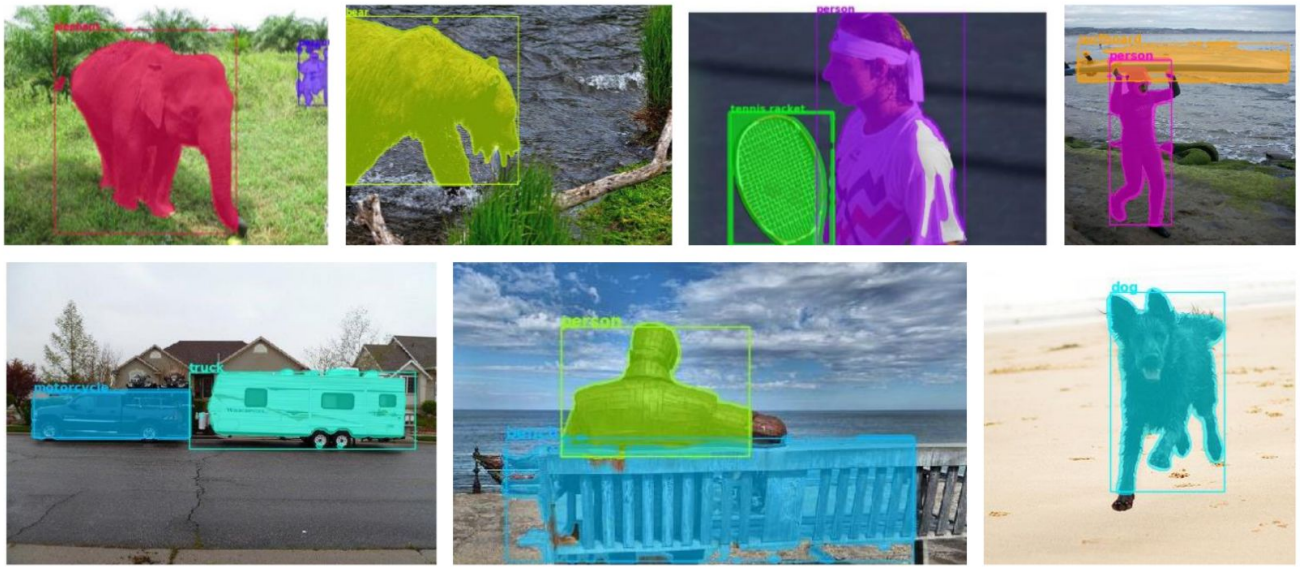


Figure 11. Visualization of pseudo-mask generated for COCO dataset [30] using our pipeline. Note that, the generated box-level and pixel-level annotations are noisy (incomplete mask and less accurate bounding box).

VS, Vibashan, et al. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 23539-23549.

Qualitative results

Figure 12. Visualization of pseudo-mask generated for Open Images [22] dataset using our pipeline. Note that, the generated box-level and pixel-level annotations are noisy (incomplete mask and less accurate bounding box).

VS, Vibashan, et al. Mask-free ovis: Open-vocabulary instance segmentation without manual mask annotations. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023. p. 23539-23549.

Qualitative results

Figure 13. Visualization of Mask-RCNN [16] predictions trained on pseudo-masks generated on COCO and Open Images - top and bottom row, respectively. Mask-RCNN training helps the model learn to filter the noise present in the pseudo-mask producing better-quality (complete mask and tight bounding box) predictions.