Single-Stage Open-world Instance Segmentation with Cross-task Consistency Regularization

Xizhe Xue [*,d], Dongdong Yu[b], Lingqiao Liu[c], Yu Liu[b], Ying Li[†,d], Zehuan Yuan[b], Ping Song[b], and Mike Zheng Shou[d]

[a]Northwestern Polytechnical University
[b]ByteDance Inc.
[c]University of Adelaide
[d]Show Lab, National University of Singapore

Abstract

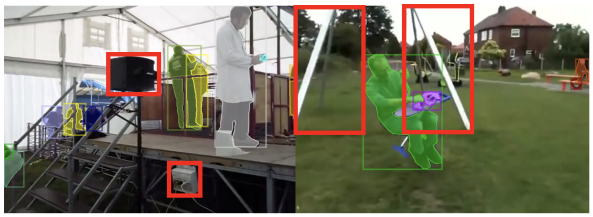
Open-world instance segmentation (OWIS) aims to segment class-agnostic instances from images, which has a wide range of real-world applications such as autonomous driving. Most existing approaches follow a two-stage pipeline: performing class-agnostic detection first and then class-specific mask segmentation. In contrast, this paper proposes a single-stage framework to produce a mask for each instance directly. Also, instance mask annotations could be noisy in the existing datasets; to overcome this issue, we introduce a new regularization loss. Specifically, we first train an extra branch to perform an auxiliary task of predicting foreground regions (i.e. regions belonging to any object instance), and then encourage the prediction from the auxiliary branch to be consistent with the predictions of the instance masks. The key insight is that such a cross-task consistency loss could act as an error-correcting mechanism to combat the errors in annotations. Further, we discover that the proposed cross-task consistency loss can be applied to images without any annotation, lending itself to a semi-supervised learning method. Through extensive experiments, we demonstrate that the proposed method can achieve impressive results in both fully-supervised and semi-supervised settings. Compared to SOTA methods, the proposed method significantly improves the AP_{100} score by 4.75% in UVO→UVO setting and 4.05% in COCO→UVO setting. In the case of semi-supervised learning, our model learned with only 30% labeled data, even outperforms its fully-supervised counterpart with 50% labeled data. The code will be released soon at: https://github.com/showlab/SOIS.

1 Introduction

Existing instance segmentation task [1] often assumes that the objects in the image can be categorized into a finite set of classes. Such an assumption, however, can be easily violated in many real-world applications, and it is very common to encounter unknown or novel objects at the inference time. The failure to consider such a scenario might cause severe secure problems for applications like autonomous driving or human-robot interaction. Therefore, the Open-World Instance Segmentation (OWIS) task, which aims to segment any object in an image, has received increasing attention recently. Several methods have been proposed to handle the OWIS task. Wang et al. [2] presented the first professional dataset UVO. OLN [3] replaces the classification branch in MaskRCNN [1] with a novel objectness branch. Saito et al. proposed the LDET [4] which mainly contains a data

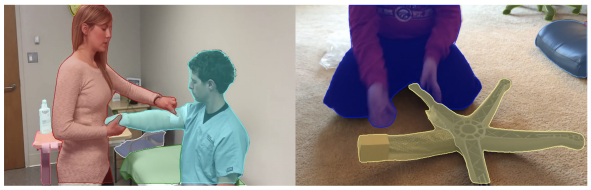
[*]This work was done when Xizhe Xue visited National University of Singapore. Email: xuexizhe@mail.nwpu.edu.cn
[†]Corresponding author

Preprint. Under review.

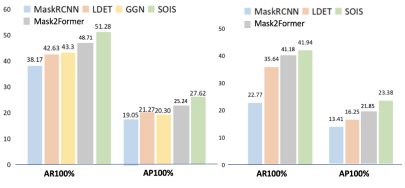


(a). Annotation samples from UVO dataset

(b). Annotation sample from COCO dataset



(c). Visualization results of our SOIS on UVO dataset



(d). COCO→UVO results (e). UVO→UVO results

Figure 1: **(a-b). Instances missing annotations in UVO and COCO datasets.** The regions in red boxes are mistakenly annotated as background. **(c). Visualization results of our SOIS on UVO dataset.** Here, the proposed SOIS is trained on COCO dataset and tested on UVO dataset. Our methods correctly localizes many objects that are not labeled in COCO with guidance of auxiliary task. **(d). The performance of our SOIS *vs.* SOTA methods on COCO→UVO. (e). The performance of our SOIS *vs.* SOTA methods on UVO→UVO.**

augmentation strategy and a decoupled training approach. GGN et al. [5] exploits additional pseudo-label supervision to learn instance boundaries and generalizes well to unseen classes.

This paper studies the OWIS task and makes two key contributions. First, unlike previous methods that adopt a two-stage detection process, we propose a **single-stage open-world instance segmentation** method by upgrading the recently proposed Mask2former framework [6]. The second and the more important contribution is a new regularization loss to guide the learning process of OWIS. Our method is motivated by the observation that the instance annotation in the existing datasets is incomplete. Unlike in closed-world instance segmentation, where the object categories have been clearly defined, instance definition in OWIS is much more ambiguous and harder for annotators to follow. Inevitably, the instance annotation could become inconsistent across images.

Our solution to this issue is to introduce a self-correcting mechanism to combat erroneous annotations. We first introduce an auxiliary task: predicting the foreground region, which should be pixels falling into *any* object instances. Then we create a prediction branch to generate the **foreground region prediction**. It is clear that the prediction target of this auxiliary branch has an inherent connection to the object instance: the former is the union of the latter. We thus regularize the foreground prediction and the instance prediction to follow the same relationship. This gives rise to a loss function dubbed **cross-task consistency loss**. Such a loss works as a self-correcting mechanism; it provides additional guidance to both prediction tasks when the noisy annotations fail to provide correct supervision.

Moreover, the proposed cross-task consistency loss could also work when no annotations are provided. This inspires us to explore its usage in a **semi-supervised setting**. In such a setting, we could learn the OWIS model from both labeled and unlabeled images. This is an attractive solution since it can significantly reduce the cost of annotating clas-agnostic instances. Specifically, we perform semi-supervised OWIS by first warming up the network on the labeled set and then continuing training it with the cross-task consistency loss on the mixed labeled and unlabeled data.

In summary, the contributions of this work can be concluded as follows.

- We design the first Single-stage Open-world Instance Segmentation framework (SOIS) without the need of an open-world proposal generation stage.
- We propose a cross-task consistency loss that can overcome the limitation of noisy instance mask annotations.
- We further extend the proposed method into a semi-supervised OWIS model, which can effectively make use of the unlabeled images to help the OWIS model training.

- Through our intensive experimental evaluation, the proposed method reaches the leading OWIS performance in the fully-supervised learning, as shown in Figure 1. Also, we demonstrate that our semi-supervised extension can achieve remarkable performance with much smaller amount of labeled data.

2 Related Work

Closed-world instance segmentation task (CWIS) [1, 7, 8, 9, 10] requires the approaches to assign a class label and instance ID to every pixel. Two-stage CWIS approaches, such as Mask RCNN, always include a bounding box estimation branch and an FCN-based mask segmentation branch. To improve efficiency, one-stage methods such as CenterMask [8], YOLACT [9], SOLO [10] and BlendMask [7] have been proposed, which remove the proposal generation and feature grouping process. In recent years, the methods [11, 12], following DETR [13], consider the instance segmentation task as an ensemble prediction problem. In addition, Cheng et al. proposed the universal segmentation framework MaskFormer [14] and Mask2Former [6], in which Mask2Foremr even outperforms state-of-the-art specialized architectures on the CWIS task.

Open-world detection and entity segmentation The open-world object detection task (OWOD) [3, 15, 16] is first proposed in [17]. It aims to detect known categories of objects and identify unknown objects, while requiring the method to have the ability to lean new classes. To address this problem, Joseph et al. [17] constructed ORE, which comprises three dedicated components namely contrastive clustering module, unknown-aware proposal network and the energy-based unknown identification module. Another method OW-DETR [15] employs the attention-driven pseudo-labeling, novelty classification, and objectness scoring jointly to meet OWOD challenges. Open-world entity segmentation task (OWES) [18] aims to segment all entities (instances and stuff) in an image without predicting their categories. To solve this problem, Qi et al. [18] proposed a center-based entity segmentation framework to improve the quality of the entity mask.

Open-world instance segmentation OWIS task [2] here focuses on the following aspects: (1) All instances (without stuff) have to be segmented; (2) Class-agnostic pixel-level results should be predicted with only instance ID and incremental learning ability is unnecessary. Several OWIS works have been developed recently. Yu et al. [19] proposed a two-stage segmentation algorithm, which decoupled the segmentation and detection modules during training and test. This algorithm achieves competitive results on the UVO dataset thanks to the abundant training data and the introduction of effective modules such as cascade RPN [20], SimOTA [21], etc. Another work in the same period named LDET [4] attempts to solve the noisy background annotation problem. Specifically, LDET first generates the background of the synthesized image by taking a small piece of background in the original image and enlarging it to the same size as the original image. The instance is then matted to the foreground of the synthesized image. The synthesized data is used only to train the mask prediction branch, and the rest of the branches are still trained with the original data. Very recently, Wang et al. proposed GGN [5], an algorithm that combines top-down and bottom-up segmentation ideas to reduce noise in background annotation and improve prediction accuracy by generating high-quality pseudo-labels.

However, the proposed SOIS is a single-stage framework, without relying on the open-wold proposal generation. Besides, we introduce a simple and effective auxiliary foreground prediction branch to guide the optimization process of the primary module. Modules for the primary OWIS task and those for auxiliary task are associated through a powerful cross-task consistency loss.

3 Methodology

3.1 Problem setting

The open-world instance segmentation aims to segment all the object instances (things) regardless their classes. This can be viewed as a special case of instance segmentation. Formally, in OWIS we aim to producing a set of binary masks, one for each class-agnostic instance. In each mask, the pixel value "1" indicates object and "0" indicates background.

3.2 Overview

As shown in Figure 4, the proposed SOIS framework consists of two modules that are designed for the primary OWIS task and the auxiliary task, respectively. The primary module can be divided into two branches. The mask prediction branch predicts the binary mask for each instance, whereas the

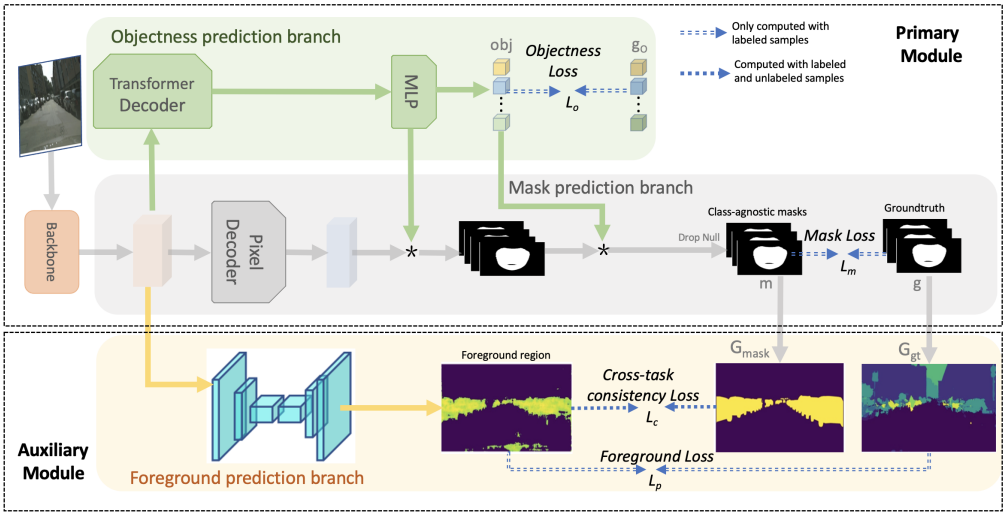


Figure 2: Overall framework of the proposed SOIS. The mask prediction branch generates the predicted masks, while the objectness prediction branch computes the objectness score for each mask. The auxiliary foreground prediction branch segments a foreground region to guide the optimization of other two branches.

objectness prediction branch estimates the weighting score for each mask. The auxiliary module contains a foreground prediction branch, which segments the foreground region, i.e., regions that belong to any object instance.

3.3 End-to-end instance prediction

Our single-stage OWIS detection is mainly achieved by the primary module. Its design is similar to the recently proposed mask2former framework [6]. Specifically, it uses the mask prediction branch to generate N binary masks with N ideally larger than the actual instance number K_i. Each mask is multiplied by a weighting score with the value between 0 and 1, indicating if a mask should be selected as an instance mask. The object score is predicted from the objectness prediction branch. Both branches can be trained by using a bipartite matching-based assignment, For more details about the training procedure, please refer to [6, 14].

In addition to the primary module, the foreground prediction branch in the auxiliary module predicts the foreground region, which is used to guide the training of the primary module through the proposed cross-task consistency loss. Once training is done, the auxiliary branch will be discarded and only the primary branch is used at the inference time.

3.4 Learning with the cross-task consistency regularization

As mentioned in the Introduction, one limitation of the OWIS study is the lack of accurate annotations due to the difficulties in annotating class-agnostic object instances. To overcome this issue, we propose a regularization to provide extra supervision to guide the OWIS model training. Such a regularization loss can be helpful when the noisy annotation fails to provide the correct guidance.

Our solution is to construct a prediction branch to predict the foreground regions whose (noisy) annotation can be derived from the (noisy) object instance annotations. Formally the foreground annotation $G_{gt}(x, y)$ is calculated by

$$G_{gt}(x,y) = \begin{cases} 0, & \text{if } \sum_{i=1}^{K} g^i(x,y) == 0 \\ 1, & \text{otherwise,} \end{cases} \tag{1}$$

where $g^i(x, y)$ is one of the K annotated object instances for the current image and the union of g^i defines the foreground object regions. Here (x, y) denotes a coordinate of a pixel in the an image. $G_{gt}(x, y)$ will be used to train the foreground prediction branch.

Our key insight is that the relationship between instance prediction and foreground prediction should follow the same relationship as in Eq 1. This relationship works as an error-correcting mechanism

and can provide supervision to models when the ground-truth supervision becomes inaccurate or unavailable. In practice, the conversion from $g^i(x, y)$ to G_{gt} is not differentiable in Eq 1. We thus approximate it by using the following equation to generate the estimate of foreground from the instance prediction:

$$G_{mask}(x, y) = \Phi\left(\sum_{j=1}^{K} m^j(x, y)\right), \tag{2}$$

where m^j means the confidence of pixels in j-th predicted mask. Φ represents the Sigmoid function.

Let the foreground prediction from the foreground prediction branch be H, our cross-task consistency loss requires the prediction from the foreground prediction branch to be consistent with $G_{mask}(x, y)$. This leads to the following loss function:

$$L_c = \text{Dice}_{(G_{mask},H)} + \text{BCE}_{(G_{mask},H)} \tag{3}$$

where Dice and BCE denote the dice-coefficient loss [22] and binary cross-entropy loss, respectively. The detailed design of the foreground prediction branch can refer to the supplementary.

3.5 Fully-supervised learning process

The overall fully-supervised optimization of proposed SOIS is carried out by minimizing the following joint loss formulation L_f,

$$L_f = \alpha L_m + \beta L_p + \gamma L_c + \omega L_o \tag{4}$$
$$L_m = \text{Dice}_{(m,g)} + \text{BCE}_{(m,g)} \tag{5}$$
$$L_p = \text{Dice}_{(H,G_{gt})} + \text{BCE}_{(H,G_{gt})} \tag{6}$$
$$L_o = \text{BCE}_{(obj,g_o)} \tag{7}$$

where L_m, L_p and L_o denote the loss terms for mask prediction, auxiliary foreground prediction and objectness scoring, respectively. α, β, γ and ω are the weights of the corresponding losses. g_0 is a set of binary value that indicate whether each mask is an instance. Before computing the L_m, matching between the set of predicted masks and groundtruth has been done via the bipartite matching algorithm.

3.6 Extension to semi-supervised OWIS learning

Due to the ambiguity of the instance definition in OWIS, it is much harder for the annotators to follow the annotation instruction, and this could make the annotations for OWIS expensive. It is desirable that one can use unlabeled data to help OWIS model training.

Our proposed cross-task consistency loss only requires the prediction from both predictors to follow an inherent relationship, and there is no ground-truth supervision needed. Thus, it is possible to apply this loss to unlabeled data. In this work, we extend the proposed method to a semi-supervised setting, in which the OWIS model can be learned from both labeled and unlabeled data, achieving a good trade-off between the annotation cost and model accuracy.

Semi-supervised learning process Given a labeled set $D_l = \{(x_i, y_i)\}_{i=1}^{N_l}$ and an unlabeled set $D_u = \{x_i\}_{i=1}^{N_u}$, our goal is to train an OWIS model by leveraging both a large amount of unlabeled data and a smaller set of labeled data. Specifically, D_l is first utilized to train the SOIS in a warm-up stage, giving an good initialization for the following semi-supervised learning. After that, we jointly train the OWIS model on the labeled data and unlabeled data. For labeled data, we employ the loss function defined in Eq 4. For unlabeled data, we only apply the cross-task consistency loss L_c.

4 Experiments

For demonstrating the effectiveness of our proposed SOIS framework, we compared it with some advanced fully-supervised methods through intra-dataset and cross-dataset evaluations. Ablation studies have also been performed in these two settings to show the effect of each component. As the proposed cross-task consistency loss extends SOIS to have the ability of semi-supervised learning, we also train SOIS in the semi-supervised setting and test it on the UVO validation subset.

Table 1: State-of-the-art comparison for SOIS on UVO dataset. All methods have been trained with the UVO-train subset and tested on the UVO-val subset.

Metric	Backbone	$AP_{100}(\%)$	$AP_s(\%)$	$AP_m(\%)$	$AP_l(\%)$	$AR_{100}(\%)$	$AR_{10}(\%)$
MaskRCNN	R-50	13.41	4.91	12.33	17.45	22.77	20.01
LDET	R-50	16.25	3.27	13.58	22.93	35.64	23.73
Mask2Former	R-50	21.85	6.16	16.82	31.65	41.18	28.26
SOIS (Ours)	R-50	**23.38**	**6.59**	**17.35**	**34.23**	**41.94**	**29.24**
Mask2Former	Swin-B	33.27	9.34	25.21	47.80	50.81	37.49
SOIS (Ours)	Swin-B	**38.02**	**12.31**	**28.64**	**53.22**	**54.74**	**41.78**

Table 2: Results of Mask2Former and proposed SOIS on COCO2017-val (none-VOC) and COCO2017-val. The models have been trained on COCO2017-train (VOC). Even though the small number of the classes of training data somewhat limits the model's ability to learn generic instance representations, the proposed SOIS still outperforms the baseline on $AR_{100}(\%)$.

Test Dataset	COCO (NoneVOC)				COCO
Metrics	$AR_{100}(\%)$	$AR_s(\%)$	$AR_m(\%)$	$AR_l(\%)$	$AR_{100}(\%)$
Mask2Former	9.21	4.56	8.79	19.30	37.22
SOIS	11.03	4.87	9.24	26.81	38.34

4.1 Implementation details and evaluation metrics

Implementation details Detectron2 [23] is used to implement the proposed SOIS framework, multi-scale feature maps are extracted from the ResNet-50 [24] or Swin Transformer [25] model pre-trained on ImageNet [26]. Our transformer encoder-decoder designs follow the same architecture as in Mask2Former [6]. The number of object queries M is set to 100. Both the ResNet and Swin backbones use an initial learning rate of 0.0001 and a weight decay of 0.05. A simple data augmentation method, Cutout [27], is applied to the training data. All the experiments have been done on 8 NVIDIA V100 GPU cards with 32G memory.

Pseudo-labelling for COCO train set Pseudo-labeling is a common way to handle noisy labels. To explore the compatibility of proposed SOIS and the pseudo-labeling operation, we further introduce a simple strategy to generate the pseudo-labels for unannotated instances in the COCO train set [28] in the following experiments.

Specifically, we follow a typical self-training framework, introducing the teacher model and student model framework to generate pseudo-labels. These two models have the same architecture, as shown in Figure 4, but are different in model weights. The weights of the student model are optimized by the common back-propagation, while the weight of the teacher model is updated by computing the exponential moving averages (EMA) of the student model. During training, the image i is first fed into the teacher model to generate some mask predictions. The prediction whose confidence is higher than a certain value would be taken as a pseudo-proposal. The state S_{ij} of the pseudo-proposal p_{ij} is determined according to Equation (8).

$$S_{ij} = \begin{cases} \text{True}, & \text{if } \operatorname{argmax}(\varphi(p_{ij}, g_i)) \leqslant \varepsilon, \\ \text{False}, & \text{otherwise}, \end{cases} \tag{8}$$

in which g_i means any instance groundtruth in the image i. φ denotes the IOU calculating function and ε is a threshold to further filter the unreliable pseudo-proposals. Finally, pseudo-proposals with states $True$ would be considered as reliable pseudo-labels. Here, the confidence and IOU threshold ε for selecting pseudo-labels are set to 0.8 and 0.2, respectively.

Then, we jointly use the groundtruth and the pseudo-labels to form the training data annotations. If a region is identified as belonging to an instance in the pseudo-label, it will be considered as a positive sample during training.

Evaluation metrics The Mean Average Recall (AR) and Mean Average Precision (AP) [28] are utilized to measure the performance of approaches in a class-agnostic way.

Table 3: State-of-the-art comparison for SOIS on UVO and LVIS dataset. All methods have been trained on the COCO dataset, while tested on the UVO-val subset and the validation set of long-tailed dataset LVIS. Our SOIS achieves favorable performance on both UVO and LVIS.

Test dataset	UVO					LVIS				
Metric	AR_{100} (%)	AP_{100} (%)	AP_s (%)	AP_m (%)	AP_l (%)	AR_{100} (%)	AP_{100} (%)	AP_s (%)	AP_m (%)	AP_l (%)
MaskRCNN-R50	38.17	19.05	6.27	13.15	28.05	22.36	6.46	3.23	10.39	17.64
LDET-R50	42.63	21.27	5.66	17.52	30.08	25.10	6.47	2.78	10.57	18.38
GGN-R50	43.30	20.30	**8.70**	18.20	27.30	–	–	–	–	–
SOLO V2-R50	39.41	22.25	5.56	14.18	34.12	21.68	7.54	3.05	12.42	22.39
SOLO V2-R50+Ours	42.52	25.04	6.77	16.90	38.33	22.81	7.86	3.79	13.06	23.75
Mask2Former-R50	48.71	25.24	6.46	16.09	40.37	24.50	8.13	2.93	13.60	26.26
SOIS-R50 (Ours)	**51.28**	**27.62**	7.80	**18.61**	**43.42**	**25.17**	**8.50**	**3.41**	**13.81**	**26.43**
Mask2Former-SwinB	51.38	28.16	7.29	18.91	45.48	25.34	9.63	4.32	**15.28**	29.02
SOIS-SwinB (Ours)	**54.86**	**32.21**	**9.03**	**21.92**	**50.69**	**25.40**	**9.88**	**4.92**	15.03	**30.08**

4.2 Fully-supervised experimental setting

Intra-dataset evaluation UVO is the largest open-world instance segmentation dataset. Its training and test images are selected from the same domain, while they do not have any overlap. Here, we perform the leaning process of SOIS on the UVO-train subset and conduct the test experiments on the UVO-val subset. Besides, we split the COCO dataset into 20 seen (VOC) classes and 60 unseen (none-VOC) classes. We train a model only on the annotation of 20 VOC classes and test it on the 60 none-VOC class, evaluating its ability of discovering novel objects.

Cross-dataset evaluation Open-world setting assumes that the instance can be novel classes in the target domain. Therefore, it is essential for the OWIS method to handle the potential domain gap with excellent generalization ability. Cross-dataset evaluation, in which training and test data come from different domains, is necessary to be conducted. Here, we first train the proposed SOIS model and compared methods on the COCO-train subset, while testing them on the UVO-val and LVIS-val [29] datasets to evaluate their generalizability. Then we extend the experiments to an autonomous driving scenario, training the models on the Cityscapes [30] dataset and evaluating them on the Mapillary [31].

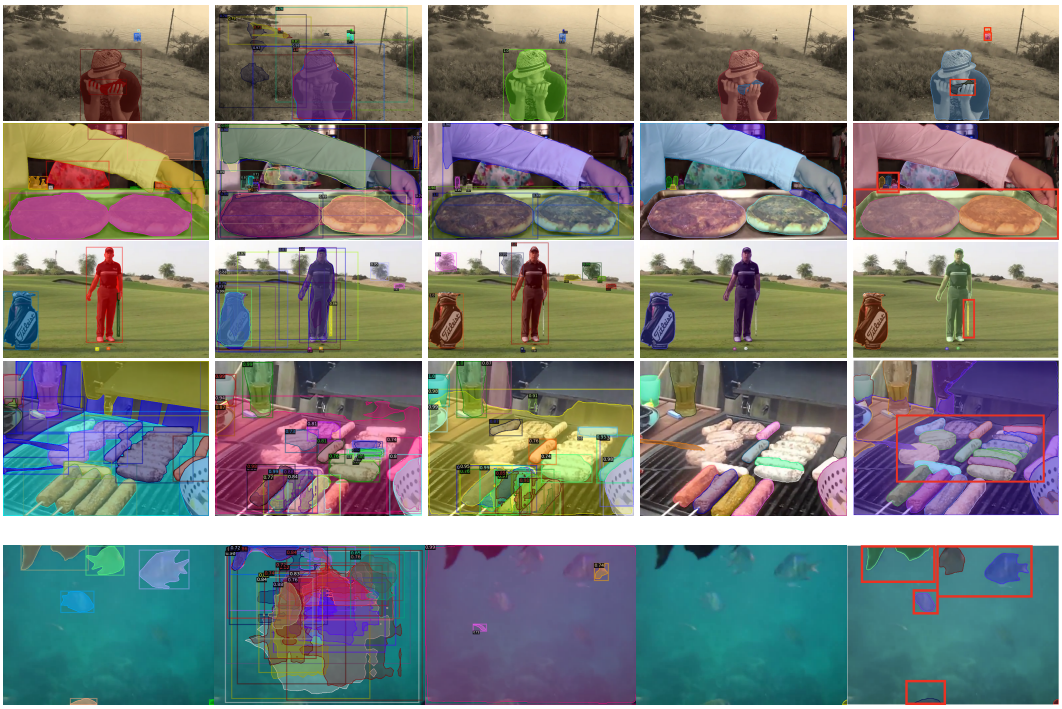
Table 4: Experiments on autonomous driving scenes. Models are trained on Cityscape (8 foreground classes, person, rider, car, truck, bus, train, motorcycle, and bicycles) and tested on the validation set of Mapillary Vistas with 35 foreground classes including not only vehicles, but also animals, trash can, mailbox, and so on.

Method	MaskRCNN	LDET	Mask2Former	OSIS(Ours)
AP(%)	7.3	7.8	7.6	8.4
AR_{10}(%)	6.1	5.5	7.0	7.5

4.3 Fully-supervised experimental results

Intra-dataset evaluation The results are illustrated in Table 1. The single-stage approaches based on the mask classification framework perform better than other two-stage methods. Among then, our proposed SOIS achieves a significant performance improvement over the baseline Mask2Former, which is 4.75% in AP_{100} and 3.93% in AR_{100} when using the Swin-B backbone. For VOC→none-VOC setting, the experimental results are shown in Table 2, which verified that our proposed method can improve the performance for all size of instances, especially for the large instances.

Cross-dataset evaluation For the COCO→UVO task, according to Table 3, it is clear that the proposed SOIS outperforms all previous methods, achieving a new state-of-the-art AR_{100} at 54.86% which is 11.56% higher than previous state-of-the-art method GGN [5]. We also applied the proposed techniques on the classic one-stage method SOLO V2 [32]. The experimental results show that it improves AR_{100} and AP_{100} by 3.11% and 2.79% compared to the SOLO V2. For the COCO→LVIS task, the overall AP and AR of SOIS still surpass the performance of other state-of-the-art methods,



(a) Groundtruth (b) MaskRCNN (c) LDET (d) Mask2Former (e) SOIS (Ours)

Figure 3: Visualization results of COCO→UVO cross-dataset evaluation. The predicted boxes of two-stage methods MaskRCNN and LDET are also drawn. Compared with other three approaches, our SOIS shows superior ability in discovering novel objects, as shown in regions in red boxes.

which demonstrates the effectiveness of our proposed techniques. We further show some of the COCO→UVO visualization results in Figure 6 to qualitatively demonstrate the superiority of our method. Please refer to the supplementary for more qualitative examples.

Note that Mask2Former does not always beat MaskRCNN-based methods. Table 4 shows a case where the MaskRCNN-based method LDET outperforms Mask2Former on AP. After using our proposed module, Mask2Former beats the MaskRCNN-based method with a clear advantage. This result has verified that our proposed method achieves the superior performance, which is not just because we chose a good baseline; the more important reason behind this is that our proposed idea does work efficiently.

4.4 Ablation study

We perform cross-dataset and intra-dataset ablation studies to analyze the effectiveness of each component in the proposed SOIS, including the auxiliary foreground prediction branch and the cross-task consistency loss. Besides, the pseudo-label generation strategy applied to handle the background annotation noise is also taken into consideration. Using the SwinB backbone, these models are trained on the COCO-train subset and the UVO-train subset, respectively. The metrics reported in Table 5 are tested on the UVO-val dataset.

Effectiveness foreground prediction branch Table 5 shows that although a separate foreground prediction branch can guide the method to optimize towards the direction of discovering foreground pixels, it only slightly boosts the performance.

Effectiveness of cross-task consistency loss Cross-task consistency loss has a positive effect on both sparse annotated (COCO) and dense annotated (UVO) training dataset. The values of AP_{100} and AR_{100} increase significantly (2.74% ↑ and 2.49% ↑ on COCO while 2.90% ↑ and 3.35%↑ on UVO) after applying the cross-task consistency loss as well as the foreground prediction branches together. This result outperforms the SOIS countpart with only pseudo-labling, showing our effectiveness. Besides, jointly utilizing our cross-task consistency loss as well as the pseudo-labling strategy leads to performance improvements on two setting, which reflects the compatibility of our method.

Table 5: Ablation results of the proposed components by cross-dataset and intra-dataset evaluations.

Component			Train on COCO		Train on UVO	
Foreground prediction	Cross-task consistency loss	Pseudo label	AP_{100}(%)	AR_{100}(%)	AP_{100}(%)	AR_{100}(%)
			28.65	51.54	35.12	51.39
✓			29.02	51.60	35.55	51.73
		✓	30.09	52.97	32.94	51.64
✓	✓		31.39	53.83	**38.02**	**54.74**
✓		✓	30.17	52.98	33.35	50.90
✓	✓	✓	**32.21**	**54.86**	37.71	52.27

Table 6: Results of models trained with only labeled data (LDET, Mask2Former, Fully-SOIS) and trained with both labeled and unlabeled data(Mean Teacher, Pseudo Labeling, Semi-SOIS) on UVO-val dataset.

Training Data	30% labeled data in UVO				50% labeled data in UVO			
Method	Fully-SOIS$_{30}$	Mean Teacher	Pseudo Labeling	Semi-SOIS$_{30}$	LDET$_{50}$	Mask2Former$_{50}$	Fully-SOIS$_{50}$	Semi-SOIS$_{50}$
AP_{100}(%)	21.67	21.95	22.77	25.03	10.61	19.49	22.86	25.22
AR_{100}(%)	40.09	40.82	41.56	45.42	25.08	38.08	41.44	47.56

Effectiveness of pseudo-labeling Pseudo-labeling is not always necessary and powerful for any types of datasets. As shown in Table 5, the AP_{100} and AR_{100} of the COCO trained model increase by 1.35% and 0.78%, respectively, after applying the pseudo-label generation. However, pseudo-labeling causes a performance degradation (e.g. 2.18%↓ in AP_{100}) to a model trained in the UVO dataset. Compared with COCO, the UVO dataset is annotated more densely. We conjecture that the background annotations of UVO are more reliable than those of COCO, where carefully selected pseudo-labels are more likely to represent unlabeled objects. The generated pseudo-labels of UVO contain higher noises than those of COCO. These additional noisy labels mislead the model training.

4.5 Semi-supervised learning experiment

Experimental setting We have divided the UVO-train dataset into the labeled subset D_l and the unlabeled subset D_u. Semi-supervised model Semi-SOIS is optimized as described in Section 3.6 on $D_l \cup D_U$, while the fully-supervised method Fully-SOIS is trained merely on the D_L. To ensure the comprehensiveness of the experiment, two different data division settings are included in our experiments:{D_L=30%, D_u=70%} and {D_L=50%, D_u=50%}. The backbone applied here is SwinB. We also implemented the classic Mean teacher model and a simple pseudo-label method based on the Mask2Former to perform comparison.

Results and analysis As presented in Table 6, the Semi-$SOIS_{50}$ model trained on the UVO with 50% annotated data outperforms the Semi-$SOIS_{30}$ model leaning with 30% labeled training images. However, the performance increase between the Semi-$SOIS_{30}$ and Semi-$SOIS_{50}$ is slight. In addition, Semi-$SOIS_{30}$ improves Fully-$SOIS_{30}$ by 3.36% and 5.33% in AP_{100} and AR_{100}, respectively. Compared to Fully-$SOIS_{50}$, Semi-$SOIS_{50}$ still achieves significant advantages (2.36% in AP_{100} and 6.12% in AR_{100}). These results reflect that cross-task consistency loss has the ability to extract information from unlabeled data and facilitates model optimization in the semi-supervised setting. It is notable that the results of Semi-$SOIS_{30}$ are even better than those of Fully-$SOIS_{50}$. This illustrates that the information dug out by the cross-task consistency loss from the remaining 70% unlabeled data is more abundant than that included in 20% fully-labeled data. Therefore, our algorithm can achieve better performance with fewer annotations. This characteristic is promising in solving the OWIS problem, where labeling is very expensive.

5 Conclusion

This paper proposes the first single-stage framework (SOIS) for the open-world instance segmentation task. Apart from predicting the instance mask and objectness score, our framework introduces an auxiliary foreground prediction branch to segment the regions belonging to any instance. In addition, a cross-task consistency loss is proposed to enforce the foreground prediction to be consistent with the prediction of the instance mask, acting as an error-correcting mechanism for handling noisy annotations. Extensive experiments demonstrate that SOIS outperforms state-of-the-art methods by a large margin on typical datasets. Further, the novel cross-task consistency loss also allows SOIS to

be optimized by using the labeled and unlabeled data jointly in a semi-supervised principle. It is an important step toward reducing expensive laborious human supervision.

Limitation and social impact: The proposed SOIS shows relatively strong generalizability through cross-dataset evaluation. However, further improvement could be made when there are conflicting foreground annotations between the training and the test dataset. For example, COCO and LVIS may have different granularities of annotation for the same object. In this case, it is worthwhile studying how to make the model unify the level of annotation. The development of OWIS methods will promote many applications in the relative industry, such as hand manipulation or robotic navigation.

Acknowledgments and Disclosure of Funding

Mike Zheng Shou is supported only by the National Research Foundation, Singapore under its NRFF award NRF-NRFF13-2021-0008.

A Appendix

In this appendix, we provide the architecture of the foreground prediction branch (in Figure 4) and detailed experimental settings first. Then some annotations in UVO dataset are visualized in Figure 5 to show the challenges of open world instance segmentation. Finally, additional visualization results of proposed SOIS are shown in Figure 6 and **??**.

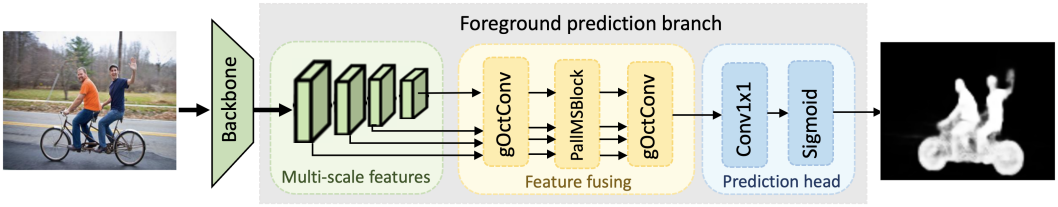


Figure 4: **Architecture of foreground prediction branch**. Multi-scale features extracted from backbone are fed into the feature fusing module to exchange and fuse the multi-scale information. Then a fused feature is sent to the prediction head to predict the final foreground map. Considering the efficiency, we follow [33] to introduce the gOctConv [33] and PallMSBlock [33] to perform feature fusing.

A.1 Detailed experimental settings

Implementation details For feature extracting, we obtain the multi-scale features through a sequential backbone network [24, 25], and FPN [34]. The multi-scale features contain D-dimensional feature maps with resolutions of 1/4, 1/8, 1/16, and 1/32. In the pixel decoder module, six MSDeformAttn layers are employed, while the transformer decoder have three layers with 100 queries by default.

In fully-supervised learning, the total loss L_f can be formulated as: $L_f = \alpha L_m + \beta L_p + \gamma L_c + \omega L_o$. We set the weight α of mask loss (L_m) to 5.0, the weight β of foreground loss (L_p) to 1.0, the weight γ of cross-task consistency loss (L_c) to 1.0 and the weight ω of objectness loss (L_o) to 2.0.

Training settings Specifically, AdamW [35] optimizer and the step learning rate schedule are applied to optimize our model. An initial learning rate of 0.0001 and a weight decay of 0.05 are utilized for all backbones. We set a learning rate multiplier of the backbone to 0.1 and we decay the learning rate at 0.9 and 0.95 fractions of the total number of training steps by a factor of 10. For data augmentation, we use the large-scale jittering (LSJ) augmentation with a random scale sampled from range 0.1 to 2.0 followed by a fixed size crop to 1024×1024 on COCO dataset and 640×640 on UVO dataset. Besides, a Cutout [36] strategy that randomly cuts out a region of size [1/8·w, 1/8·h] to [1/3·w, 1/3·h] is introduced during training. On COCO dataset, we train our models for 38×10^4 iterations with a batch size of 16, while on UVO dataset, we train our models for 12×10^4 iterations with the same batch size.

SOIS training process with pseudo-labeling on COCO dataset

Algorithm 1: SOIS training process with pseudo-labeling

Data: Image dataset
Result: Proposed SOIS Model M_u
initialization the student model M_u, and teacher model M_t=M_u.copy();
while *Image* $i \notin \varnothing$ **do**
  read image i and corresponding groundtruth gt_i;
  extract backbone feature X_i;
  pred_masks ← M_t.predictor(X_i);
  pseudo_proposals← filter_masks_with_confidence(pred_masks, confidence_threshold);
  pseudo labels ← filter_masks_with_IOU(pseudo_proposals, IOU_threshold);
  training labels ← merge(gt_i, pseudo labels);
  aug_data← Cutout(X_i, training labels);
  M_u ← M_u.training(aug_data);
  M_t ← M_t.EMA_update(M_t,M_u)
end

Figure 5: Visualizations of UVO annotations. It is notable that the same class of object may be labeled as an instance or as background in different images. (as shown in the area highlighted by the ellipse). This inconsistency of annotations pose a great challenge to the algorithms.

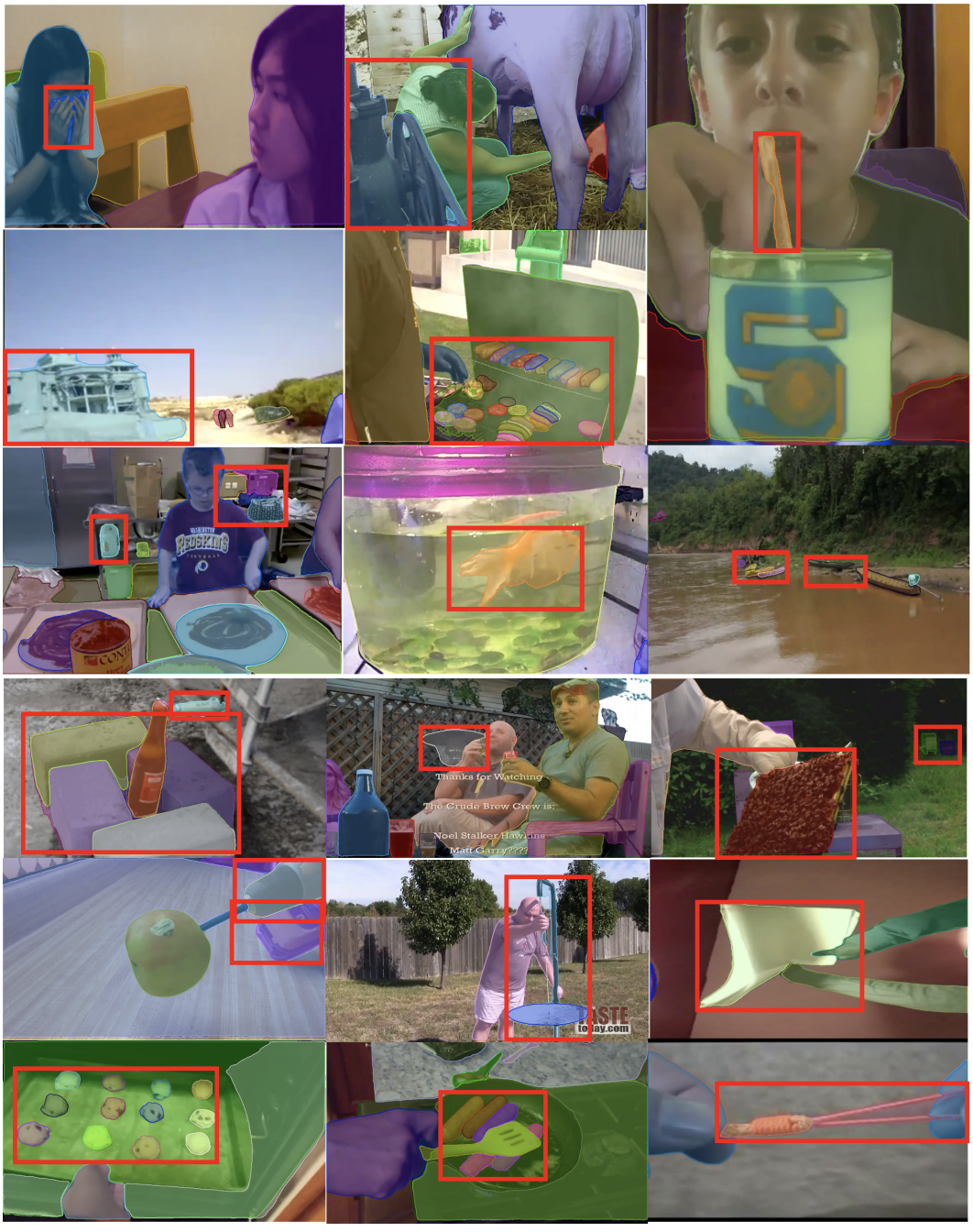


Figure 6: Visualizations results of our proposed SOIS in UVO dataset. SOIS can discover many novel objects, as shown in regions in red boxes.

A.2 Visualization of annotations and our results on UVO dataset

Unlike in closed-world instance segmentation, where the object categories have been clearly defined, instance definition in OWIS is much more ambiguous and harder for annotators to follow. Inevitably, the instance annotation could become inconsistent across images, as shown in Figure 5. Our method is motivated by this observation that the instance annotation in the existing datasets is very noisy. Our solution to this issue is to introduce a self-correcting mechanism to combat erroneous annotations, which provides additional guidance to both prediction tasks when the noisy annotations fail to provide correct supervision. The visualization results in Figure 6 demonstrate that our proposed SOIS can segment many novel objects that have not been unseen in the training set.

References

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[2] W. Wang, M. Feiszli, H. Wang, and D. Tran, "Unidentified video objects: A benchmark for dense, open-world segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10776–10785, 2021.

[3] D. Kim, T.-Y. Lin, A. Angelova, I. S. Kweon, and W. Kuo, "Learning open-world object proposals without learning to classify," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5453–5460, 2022.

[4] K. Saito, P. Hu, T. Darrell, and K. Saenko, "Learning to detect every thing in an open world," *arXiv preprint arXiv:2112.01698*, 2021.

[5] W. Wang, M. Feiszli, H. Wang, J. Malik, and D. Tran, "Open-world instance segmentation: Exploiting pseudo ground truth from learned pairwise affinity," *CVPR*, 2022.

[6] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," 2022.

[7] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, "Blendmask: Top-down meets bottom-up for instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8573–8581, 2020.

[8] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3150–3158, 2016.

[9] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9157–9166, 2019.

[10] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "Solo: Segmenting objects by locations," in *European Conference on Computer Vision*, pp. 649–665, Springer, 2020.

[11] Y. Fang, S. Yang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, and W. Liu, "Instances as queries," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6910–6919, 2021.

[12] B. Dong, F. Zeng, T. Wang, X. Zhang, and Y. Wei, "Solq: Segmenting objects by learning queries," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[13] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020.

[14] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[15] A. Gupta, S. Narayan, K. Joseph, S. Khan, F. S. Khan, and M. Shah, "Ow-detr: Open-world detection transformer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2022.

[16] S. Konan, K. J. Liang, and L. Yin, "Extending one-stage detection with open-world proposals," *arXiv preprint arXiv:2201.02302*, 2022.

[17] K. Joseph, S. Khan, F. S. Khan, and V. N. Balasubramanian, "Towards open world object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5830–5840, 2021.

[18] L. Qi, J. Kuen, Y. Wang, J. Gu, H. Zhao, Z. Lin, P. Torr, and J. Jia, "Open-world entity segmentation," *arXiv preprint arXiv:2107.14228*, 2021.

[19] Y. Du, W. Guo, Y. Xiao, and V. Lepetit, "1st place solution for the uvo challenge on video-based open-world segmentation 2021," *arXiv preprint arXiv:2110.11661*, 2021.

[20] T. Vu, H. Jang, T. X. Pham, and C. Yoo, "Cascade rpn: Delving into high-quality region proposal network with adaptive convolution," *Advances in neural information processing systems*, vol. 32, 2019.

[21] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[22] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571, IEEE, 2016.

[23] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," *https://github.com/facebookresearch/detectron2*, 2019.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[27] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[28] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[29] A. Gupta, P. Dollar, and R. Girshick, "Lvis: A dataset for large vocabulary instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5356–5364, 2019.

[30] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[31] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE international conference on computer vision*, pp. 4990–4999, 2017.

[32] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "Solov2: Dynamic and fast instance segmentation," *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[33] S.-H. Gao, Y.-Q. Tan, M.-M. Cheng, C. Lu, Y. Chen, and S. Yan, "Highly efficient salient object detection with 100k parameters," in *European Conference on Computer Vision*, pp. 702–721, Springer, 2020.

[34] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[35] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[36] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.