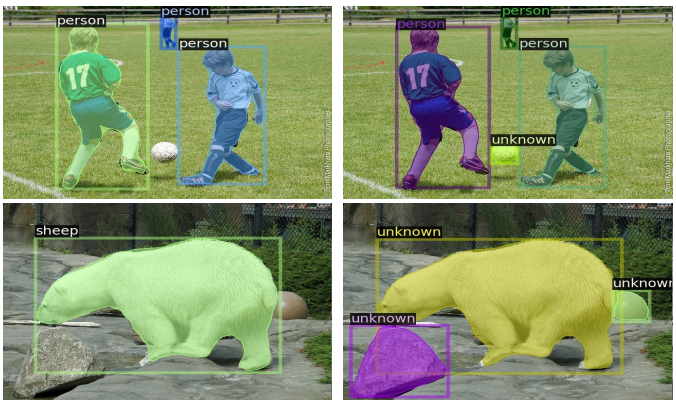
IEEE Robotics and Automation Letters (RAL) paper presented at
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Class Semantics Modulation for Open-Set Instance Segmentation

Yifei Yang[1], ZhongXiang Zhou[1], Jun Wu[1], Yue Wang[1], and Rong Xiong[1]

***Abstract*—This paper addresses the challenge of open-set instance segmentation (OSIS) which segments both known objects and unknown objects not seen in training, thus is essential for enabling robots to safely work in the real world. Existing solutions adopt class-agnostic segmentation where all classes share the same mask output layer leading to inferior performance. Motivated by the superiority of the class-specific mask prediction in close-set instance segmentation, we propose SemSeg with class semantics extraction and mask prediction modulation for conducting class-specific segmentation in OSIS. To extract class semantics for both known and unknown objects in the absence of supervision on unknown objects, we use contrastive learning to construct an embedding space where objects from each known class cluster in an independent territory and the complementary region of known classes can accommodate unknown objects. To modulate the mask prediction, we convert class semantic embedding to convolutional parameters used to predict the mask. Class semantics modulated OSIS allows optimizing the mask output layer for each class independently without competition between each other. And class semantic information is engaged in the segmentation process directly so that can guide and facilitate the segmentation task, which benefits unknown objects with severe generalization challenges particularly. Experiments on the COCO and GraspNet-1Billion datasets demonstrate the merits of our proposed method, especially the strength of instance segmentation for unknown objects.**

***Index Terms*—Visual Learning, Deep Learning for Visual Perception, Object Detection, Segmentation and Categorization.**



(a) close-set instance segmentation (b) open-set instance segmentation

(c) class-agnostic pipeline (d) class-specific pipeline

Fig. 1. (a) Close-set instance segmentation tends to ignore or mislabel unknown objects. (b) Open-set instance segmentation is asked to segment both known and unknown objects. (c) Illustration of existing OSIS solutions that adopt class-agnostic segmentation. (d) Illustration of the proposed class-specific OSIS with class semantics modulation. CLS and SEG are short for classification branch and segmentation branch respectively.

I. Introduction

In a real-world scenario, a robot will inevitably encounter objects outside its predefined taxonomy. Under the close-set assumption that all object categories present during inference are included in the training datasets, the robot often struggles to distinguish unknown objects from the background, or mislabels unknown objects as known classes as shown in Fig. 1(a), resulting in potentially catastrophic consequences. In developmental psychology [1], [2], it is believed that the ability to identify what one does not know is crucial in arousing curiosity and such curiosity fuels the desire to learn new things. Therefore, the robot needs the ability for open-set perception, not only to identify and locate known objects but also to locate objects outside training datasets and label them as unknown, as shown in Fig. 1(b).

To endow the robot with the open-set visual perception capability, open-set object detection (OSOD) [3]–[6] has recently drawn substantial attention. Compared with object detection which uses bounding boxes to locate objects, instance segmentation predicts a pixel-level mask for each object. This provides more detailed information about the object's boundaries and shape, enabling the robot to interact with objects more accurately and safely. However, limited research has focused on open-set instance segmentation (OSIS).

Intuitively, we can address the OSIS task in two ways. One is to add a class-agnostic segmentation branch after the proposal extraction network of an open-set object detector [3]–[6]. For each proposal, this method predicts a binary mask in parallel with classification. The inference results of the two branches are combined to form the final result. The other is to combine class-agnostic instance segmentation [7]–[9] with open-set recognition [10]–[12]. The class-agnostic instance segmentation model is first employed to mask all objects in an image without labels. Next, the open-set recognition model

Manuscript received: November, 1, 2023; Accepted December, 13, 2023.
This paper was recommended for publication by Editor Markus Vincze upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by the National Nature Science Foundation of China under Grant 62373322 and the National Nature Science Foundation of China under Grant 62173293. (Corresponding author: Zhongxiang Zhou.)

[1]Yifei Yang, ZhongXiang Zhou, Jun Wu, Yue Wang and Rong Xiong are with the State Key Laboratory of Industrial Control Technology and Institute of Cyber-Systems and Control, Zhejiang University, Zhejiang 310027, China yf.yang.research@outlook.com

Project page: https://yifei-y.github.io/project-pages/SemSeg/.
Digital Object Identifier (DOI): see top of this page.

classifies each object into one of the known classes or as unknown. Viewing the segmentation branch of both methods in detail, we find that the mask of an object proposal is segmented without the participation of class information, i.e., class-agnostic segmentation as shown in Fig. 1(c). In parallel, close-set instance segmentation methods [13]–[15] typically use the class-specific mask head which predicts K masks per object proposal (K is the number of classes) and selects the mask with the same index as the label predicted by the classification branch. Further, He et al. [13] found that the close-set instance segmentation model with class-specific mask head performs better than that with class-agnostic mask head. With the class-specific mask output layer, the mask prediction for each class can be optimized independently without competition between each other contributing to better segmentation quality. A natural question is: *can we apply the class-specific mask prediction to open-set instance segmentation*?

Due to the absence of unknown objects during training, implementing class-specific mask prediction in OSIS following the same way as closed-set instance segmentation, i.e., adding a specific mask output layer for unknown, is unreasonable. To this end, we propose a novel method called SemSeg that extracts and utilizes class semantics to modulate the segmentation branch, enabling class-specific mask prediction in OSIS. The basic idea is illustrated in Fig. 1(d). To extract class semantics for both known and unknown objects in the absence of supervision on unknown objects, we use contrastive learning to construct an embedding space where intra-class compactness and inter-class separation are encouraged. Distances in the learned embedding space can be a metric corresponding to semantics. And objects from each known class cluster in an independent territory in the space. The complementary region of known classes can accommodate unknown objects and the broadness of the complementary region is suitable for the diversity of unknown classes. So, the proposal embedding of known and unknown objects both contain class semantics. To utilize class semantics for mask prediction, we convert the proposal embedding to convolutional parameters used to predict the mask. In this way, SemSeg has a mask output layer modulated by class semantics for each proposal and achieves class-specific mask prediction in OSIS. Another advantage of the proposed approach is that we engage the class semantics in the segmentation process instead of selecting the segmentation result based on the predicted class. Therefore, class semantics can guide and facilitate the segmentation task, which is essential for unknown objects with severe generalization challenges. We evaluate our approach on COCO [16] and GraspNet-1Billion [17] datasets. Across all datasets, SemSeg exhibits superior instance segmentation performance over previous methods, especially on unknown objects. Our contributions can be summarized as follows:

- We propose a novel method, SemSeg, performing class-specific open-set instance segmentation with class semantics extraction and mask prediction modulation.
- Lacking unknown data for supervision, we construct an embedding space to extract semantics for both known and unknown using contrastive learning. And the class semantics are further leveraged to modulate the segmentation branch enabling class-specific mask prediction in OSIS finally.
- Extensive experimental results on two popular benchmarks demonstrate the superiority of our proposed SemSeg, especially on unknown objects.

II. Related Work

A. Open-Set Perception

The task of **open-set instance segmentation (OSIS)** is to locate all objects in an image with pixel-level masks and classify them as one of the known classes or as unknown. Below we will introduce several similar open-set perception tasks and point out the differences between them and OSIS. **Open-set recognition (OSR)** [10]–[12] aims to label the input image as one of the known classes or as unknown. While OSR methods can recognize unknown objects, but only at the image level. **Open-set object detection (OSOD)** models [3]–[6] are tasked to detect both known and unknown objects. Though researchers have made great strides in OSOD such as Bayesian SSD [3], they cannot predict masks providing more accurate localization information. **Open-set semantic segmentation (OSS)** [18]–[20] is the task of classifying pixels of an image into known and unknown classes. OSS outputs masks, but it does not distinguish different instances of the same class. Kim et al. [7] first established the protocol of **class-agnostic instance segmentation (CAIS)**, the task of localizing all objects without specifying categories. Saito et al. [8] explored a data augmentation scheme, copy-pasting known objects onto synthetic backgrounds during training, to avoid the suppression of hidden objects. However, CAIS models cannot distinguish between specific categories and consider category information as a hindrance to the generalization ability of the model rather than figuring out how to make better use of it. **Open-set panoptic segmentation (OPS)** [21], [22] requires performing panoptic segmentation for not only known classes but also unknown ones. OPS assumes that all unknown classes belong to the *thing* category and unknown objects appear and only appear in the *void* regions during training. The second assumption cannot be satisfied in many datasets, such as the datasets for robotic tabletop manipulation whose training sets contain no unlabeled unknown object. **Open-vocabulary segmentation** [23]–[25] needs to know the names of unseen classes while OSIS does not need these external knowledge. **Partially supervised instance segmentation** [26], [27] assumes that all categories have bounding box annotations during training, which is different from OSIS.

B. Contrastive Learning

Contrastive learning [28]–[30] is popularized for self-supervised learning, whose core idea lies in that positive samples are attracted while negative samples are pulled away to learn discriminative representations. Khosla et al. [31] proposed a supervised extension of contrastive learning benefiting a diversity of downstream vision tasks [32], [33]. We use supervised contrastive learning to construct a class semantics embedding space where objects from each known class cluster in an independent territory and the complementary region of

known classes can accommodate unknown objects. So we can extract class semantics of both known and unknown objects in the absence of unknown data for supervision.

C. Information Blending for Mask Prediction

The idea of blending different information for mask prediction has been widely used in close-set instance segmentation. YOLACT [34], CenterMask [35] and BlendMask [36] break up instance segmentation into two parallel tasks: generating global masks for the entire image and predicting local instance representation, which are combined to get the final mask. They focus on hybridizing global and local representation, while we attend to blending class semantics into mask prediction. CondInst [37] and follow-up [38]–[40] blend characteristics of the target instance into the segmentation branch. Our work is inspired by them, but there are twofold differences. First, we highlight the significance of class semantics and use contrastive learning to explicitly encode class semantics into the convolution filters. Second, it is non-trivial to implement class semantics modulation in OSIS. The class semantics extraction and utilization pattern needs to be tailored for generalization to unknown objects.

III. Methodology

Given an RGB image, the task of open-set instance segmentation is to locate all objects with pixel-level masks and classify them as one of the known classes or as unknown. Motivated by the superiority of the class-specific mask head in close-set instance segmentation, we propose SemSeg with class semantics extraction and mask prediction modulation for conducting class-specific segmentation in OSIS. Specifically, SemSeg constructs an embedding space containing rich class semantics and predicts masks with the participation of class semantics. The overview of the proposed model is illustrated in Fig. 2. SemSeg adopts a two-stage procedure similar to Mask R-CNN [13]. We first use ResNet [41] with Feature Pyramid Network [42] as the backbone network to extract feature maps and adopt a two-stage class agnostic detector, OLN [7], to generate object proposals. Then, segmentation and classification are applied to each proposal. Note that we do not distinguish between the first proposal generation stage and the second proposal refinement stage of OLN, but refer to them collectively as object proposal generation for clarity.

A. Object Proposals Generation

To enhance the generalization ability for unknown objects, we generate object proposals by employing OLN which learns to distinguish object proposals from the background using cues from localization quality instead of category label. The loss of object proposals generation is formulated as:

$$L_{oln} = \lambda_1 L_{ctr} + \lambda_2 L_{box_1} + \lambda_3 L_{iou} + \lambda_4 L_{box_2} \tag{1}$$

where L_{ctr}, L_{box_1}, L_{iou}, L_{box_2} are smooth L1 loss for centerness regression, lrtb bounding box regression, IoU regression and delta xywh bounding box regression. λ_1, λ_2, λ_3, λ_4 are weighting coefficients. L_{ctr} and L_{box_1} are for initial proposal generation. L_{iou} and L_{box_2} are for proposal refinement.

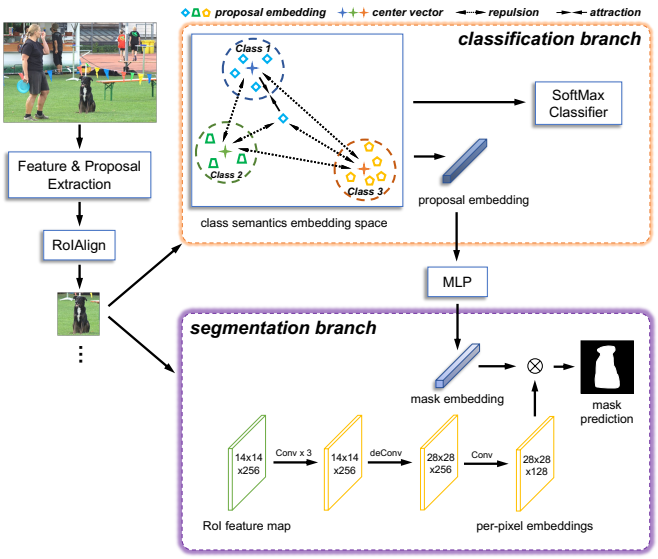


Fig. 2. Overview of SemSeg. In the classification branch, we encode each proposal into an embedding space that contains rich class semantics and can be used to differentiate between known and unknown objects. A softmax classifier then classifies known proposals into known classes. In the segmentation branch, we extract the RoI feature map for each proposal and obtain per-pixel embeddings through convolution. Finally, we predict the binary mask via a dot product ($\otimes$) between the per-pixel embeddings and the mask embedding converted from the corresponding proposal embedding.

B. Class Semantics Embedding Space

Under the OSIS setting, the training set has no labeled objects from unknown classes. To enable class-specific mask prediction in OSIS, the first problem that needs to be solved is how to extract the class semantics of unknown objects which are unavailable during training and may contain various classes. Inspired by [5], [6], [10], we turn to supervised contrastive learning for help.

Specifically, we encode proposals into an embedding space through RoIAlign [13] and three linear layers. And each known class has a learnable center vector in the embedding space. We use supervised contrastive learning to regularize the embedding space, encouraging intra-class compactness and inter-class separation. Assume we have K known classes and sample a mini-batch of N proposals, the contrastive loss is formulated as:

$$L_{cont} = \frac{1}{N}\left[\sum_{i}^{N}(L_{intra} + L_{inter}) + L_{center}\right] \tag{2}$$

$$L_{intra} = y_{ij}\max\left(D_{ij} - m_p, 0\right) \tag{3}$$

$$L_{inter} = \max_{j}\left[(1 - y_{ij})\max\left(m_n - D_{ij}, 0\right)\right] \tag{4}$$

$$L_{center} = \sum_{k}^{K}\max_{q \neq k}\left[\max\left(m_p + m_n - D^{C}_{kq}, 0\right)\right] \tag{5}$$

where D_{ij} is the cosine distance between proposal embedding z_i and center vector C_j of class j, y_{ij} is an indicating variable with value 1 if z_i is from class j and 0 otherwise, m_p and m_n are thresholds for intra-class and inter-class respectively, D^{C}_{kq} is the cosine distance between center vectors of class k and q. L_{intra} pulls the proposal and center vector of the same class

close while L_{inter} punishes the minimum distance between the proposal and center vectors of other classes. Meanwhile, L_{center} constrain the distance between center vectors.

In this way, the distance between a proposal embedding and a center vector can be a metric for the probability that the proposal belongs to the corresponding class. Therefore, each known class has a territory centered on the center vector. The complementary region of known classes can be used to represent unknown objects. And the broadness of the complementary region is suitable for the diversity of unknown classes. So proposal embedding of known and unknown objects both contain class semantics which makes it possible to leverage class semantics in OSIS.

C. Segmentation with Class Semantics Modulation

The other problem we need to address is how to give the class semantics a role in segmentation. We use class semantics to modulate the network parameters in the segmentation branch. Specifically, after RoI feature map of each proposal is extracted using RoIAlign [13], per-pixel embeddings E_{pixel} are encoded by a fully convolutional network (FCN) with ReLU in hidden layers. In the FCN, all convolutional layers are 3x3 with stride 1 and padding 1. The deconvolutional layer is 2x2 with stride 2 to increase the resolution. We convert corresponding proposal embedding to mask embedding E_{mask} by an MLP, which consists of two fully connected layers and a ReLU hidden layer. Then we predict mask logit via a dot product between the mask embedding and the per-pixel embeddings. The mask logit is followed by a sigmoid activation to get the mask prediction m for the proposal:

$$m = \text{sigmoid}(E_{mask}^T \otimes E_{pixel}) \tag{6}$$

where m and E_{pixel} are with a height and width of 28. And the channel of E_{pixel} is 128.

In training, the mask loss L_{mask} is defined only on positive proposals. The mask target is the intersection between a proposal and its associated ground-truth mask. Following [13], we define L_{mask} as the average binary cross-entropy loss.

An alternative class semantics utilization way is to blend class semantics with network input. Due to the absence of unknown objects in training, with an embedding vector of an unknown object as input, the network will face serious generalization difficulty during inference. In contrast, our modulation method does not interfere with the input to the network but with the network parameters. So, instead of struggling to accommodate the unknown class semantics, the segmentation branch can leverage the guidance of class semantics to alleviate the generalization challenge. We show the experimental comparison in Section IV-G.

The dot product is essentially the same as 1x1 convolution, where mask embedding is the convolutional kernel. Thus, our approach is equivalent to having a dedicated mask output layer modulated by class semantics for each proposal, enabling class-specific mask prediction in OSIS. With specific mask output layers for each class, the segmentation for each class can be optimized independently without competition between each other and thus can lead to better segmentation quality. In addition, engaged in the segmentation process, the class semantics can provide guidance for the segmentation task, which is especially helpful for the segmentation performance of unknown objects with severe generalization challenges, as shown in the following experiments (Section IV).

D. Open-Set Classification

The class semantics embedding space is a metric space where the cosine distance serves as the similarity metric. Therefore, known and unknown objects can be differentiated using cosine distances between proposal embedding and center vectors of known classes. If the distance between a proposal embedding and all center vectors is larger than a set threshold T_u, the proposal will be identified as unknown. Finally, a softmax classifier performs close-set classification for known proposals, which is optimized by cross-entropy loss L_{cls}.

E. Overall Optimization

SemSeg is trained in an end-to-end manner with the following multi-task loss:

$$L = \alpha L_{oln} + \beta L_{cont} + \gamma L_{cls} + L_{mask} \tag{7}$$

which is a combination of total loss of object proposals generation L_{oln}, contrastive loss L_{cont}, cross-entropy loss for softmax classifier L_{cls} and binary mask prediction loss L_{mask} with weighting coefficients α, β, γ.

IV. Experiments

To evaluate the performance of the proposed SemSeg, we first report its performance on COCO and GraspNet-1Billion datasets with comparisons to several baselines. Then we conduct deep analysis of the main components of SemSeg.

A. Comparison Methods

Since there are few studies focusing on OSIS, we implement two baselines by extending representative methods for other open-set vision tasks with intuitive adjustments:

1) OpenDet+CA-Mask: One baseline adds a class-agnostic segmentation branch after the RPN of the open-set object detector OpenDet [5]. For each proposal generated by the RPN, this model performs classification and segmentation in parallel before combining the inference results of the two branches to form the final result.

2) OLN-Mask+PROSER: The other baseline first employs the class-agnostic instance segmentation model OLN-Mask [7] to mask all objects in an image without labels. Next, the open-set recognition model PROSER [12] is used to classify each object into one of the known classes or as unknown.

In addition, we use the close-set instance segmentation model Mask R-CNN [13] as a reference to provide a more intuitive picture of the performance.

B. Implementation Details

All methods use ResNet-50 with FPN as the backbone and are trained on one NVIDIA GeForce RTX 3090 with a batch size of 4 for 128k iterations using the Detectron2 [43] framework. The SGD optimizer is adopted with an initial learning rate of 0.005 and a weight decay of 0.0001. The learning rate is decreased at the 84k and 116k iterations. At test

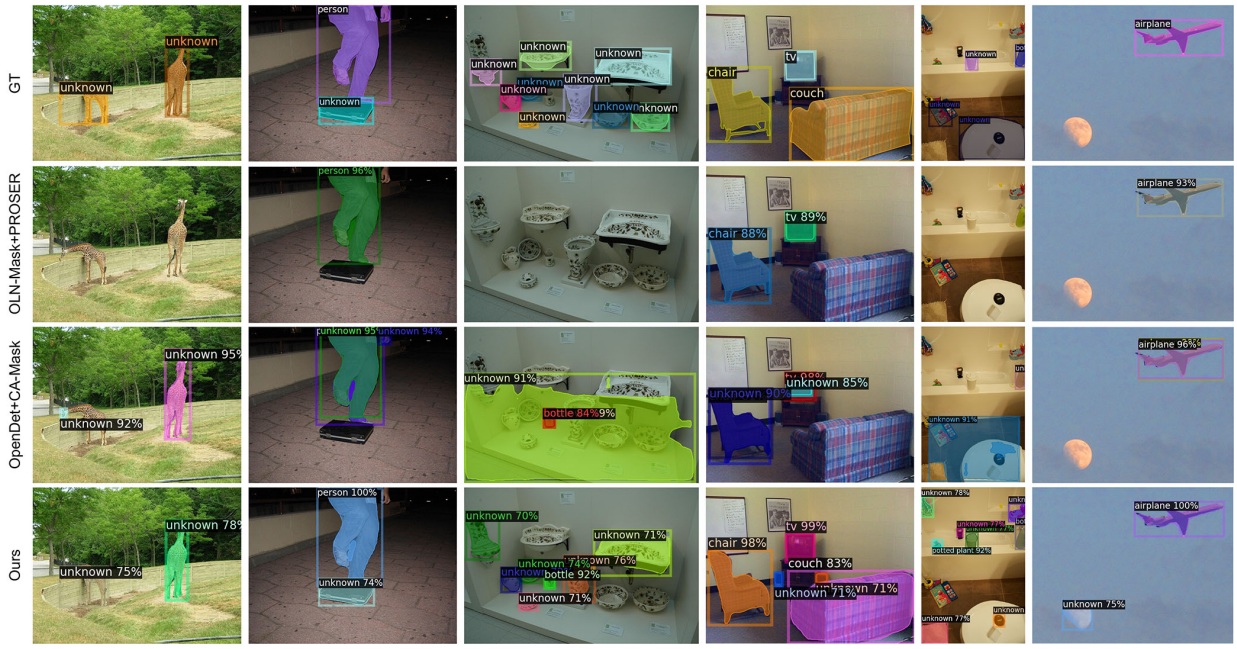


Fig. 3. Qualitative comparisons between SemSeg and other methods on COCO. Known objects with confidence greater than 0.8 and unknown objects with confidence greater than 0.7 are visualized. Note that the sound boxes, camera flash, toys, and moon in the last three columns are not labeled in the COCO dataset, but our proposed SemSeg method is able to segment them.

time, all methods keep at most 100 segmentation results. For SemSeg, we set $m_p = 0.05$, $m_n = 0.95$, $T_u = 0.085$ for the class semantics embedding space. The weighting coefficients of (1) are $\lambda_{1,2,3,4} = \{1, 10, 1, 2\}$. The loss weights in (7) are $\alpha = 1$, $\beta = 2$, $\gamma = 1$ on the COCO dataset and $\alpha = 1$, $\beta = 3$, $\gamma = 2$ on the GraspNet-1Billion dataset. The hyperparameters of the backbone network and OLN follow [6]. We also perform hyperparameter tuning for the baseline models to achieve better performance, while Mask R-CNN follows the hyperparameter setting in Detectron2.

C. Experiments on COCO

Following [7]–[9], we split the COCO2017 [16] dataset into 20 known classes and 60 unknown classes. We delete the annotations of unknown classes and filter out the images without any annotation in the training set. We evaluate models on the COCO 2017 validation set including both known and unknown classes.

For metrics, we use the Average Precision (AP) to quantify the segmentation performance of known classes (AP_k). Since the COCO dataset is not fully annotated, segmentation of unlabeled objects will be considered as False Positive wrongly, causing AP to underestimate performance on unknown objects. Therefore, we use the Average Recall (AR) as the performance metric of unknown objects (AR_{unk}). AR is closely related to the number of segmentations retained, so we report AR at different budgets (eg. AR^{10}_{unk} is the AR_{unk} in the case of retaining up to 10 unknown objects). In addition, we use Absolute Open-Set Error (AOSE) [4] to count the number of unknown objects that get wrongly classified as any of the known classes. Unless noted otherwise, all metrics are calculated using mask IoU.

Quantitative results are summarized in Tab. I. For known objects (quantified via AP_k), our model performs the best of the three OSIS models. The comparison with Mask R-CNN indicates that our model narrows the performance gap between open-set and closed-set instance segmentation tasks on known objects. As for unknown objects, SemSeg shows significant improvement over the baseline models on AR_{unk} at all budgets. And lower AOSE shows that our approach can better alleviate the problem of overconfidence in unknown objects. The qualitative results in Fig. 3 also demonstrate that SemSeg is superior to the two baseline methods. And we can get a glimpse of the segmentation performance of OSIS models on unlabeled objects in COCO. In the last three columns, SemSeg can segment some unlabeled objects including sound boxes, camera flash, toys, moon, etc.

TABLE I
COMPARISON RESULTS ON COCO AND THE BEST RESULTS AMONG OSIS MODELS ARE HIGHLIGHTED IN **BOLD**.

	AOSE ↓	AP_k ↑	AR^{10}_{unk} ↑	AR^{30}_{unk} ↑	AR^{100}_{unk} ↑
Mask R-CNN	3474	34.9	0.0	0.0	0.0
OLN-Mask + PROSER	2604	23.9	6.8	7.9	8.0
OpenDet + CA-Mask	3181	29.8	5.9	6.5	6.5
SemSeg	**2536**	**30.1**	**11.7**	**16.7**	**19.0**

D. Experiments on GraspNet-1Billion

Due to non-exhaustive annotation, the instance segmentation performance of unknown objects cannot be unbiasedly evaluated on the COCO dataset. Therefore, we conduct experiments on the fully annotated GraspNet-1Billion [17] dataset. Following the approach to construct OSOD benchmark in [6], we introduce the GraspNet OSIS benchmark by reorganizing GraspNet-1Billion. The GraspNet OSIS benchmark contains 88 classes, 28 of which are used as known and the others are unknown. The training set has 9728 images with only known classes. There are two test settings on the benchmark:

TABLE II
COMPARISON RESULTS ON GRASPNET-OSIS-T1. FOR SETTING T1, UNKNOWN CLASSES ARE GRADUALLY INCREASED IN GRASPNET-TEST-{1, 2, 3}, CONTAINING 28 KNOWN AND {12, 34, 60} UNKNOWN CLASSES RESPECTIVELY. THE BEST RESULTS AMONG OSIS MODELS ARE HIGHLIGHTED IN **BOLD**.

	GraspNet-Test-1			GraspNet-Test-2			GraspNet-Test-3		
	AOSE ↓	AP_k ↑	AP_{unk} ↑	AOSE ↓	AP_k ↑	AP_{unk} ↑	AOSE ↓	AP_k ↑	AP_{unk} ↑
Mask R-CNN	132192	65.0	0.0	310951	61.9	0.0	440288	60.8	0.0
OLN-Mask + PROSER	20961	59.6	33.2	65491	55.9	39.1	94131	54.9	39.4
OpenDet + CA-Mask	100225	63.4	21.4	237535	58.7	31.9	331990	57.4	32.0
SemSeg	**17598**	**63.6**	**37.2**	**58111**	**60.7**	**41.7**	**80546**	**60.1**	**42.2**

TABLE III
COMPARISON RESULTS ON GRASPNET-OSIS-T2. FOR SETTING T2, WILDERNESS RATIO (WR) IS GRADUALLY INCREASED IN GRASPNET-TEST-{4, 5, 6} WITH WR = {1, 2, 3} SEPARATELY. THE BEST RESULTS AMONG OSIS MODELS ARE HIGHLIGHTED IN **BOLD**.

	GraspNet-Test-4			GraspNet-Test-5			GraspNet-Test-6		
	AOSE ↓	AP_k ↑	AP_{unk} ↑	AOSE ↓	AP_k ↑	AP_{unk} ↑	AOSE ↓	AP_k ↑	AP_{unk} ↑
Mask R-CNN	52665	64.9	0.0	169551	61.0	0.0	257926	59.0	0.0
OLN-Mask + PROSER	7999	61.0	31.4	36362	55.9	40.6	56067	54.4	39.5
OpenDet + CA-Mask	39265	63.9	21.4	127570	57.1	35.4	192734	54.9	33.5
SemSeg	**6971**	**64.3**	**37.4**	**33213**	**59.9**	**44.0**	**48885**	**59.1**	**43.0**

GraspNet-OSIS-T1,T2. For setting T1, unknown classes are gradually increased to build three joint datasets: GraspNet-Test-{1, 2, 3}, containing {23808, 31744, 38912} images of 28 known classes and {12, 34, 60} unknown classes. [10] defines Wilderness Ratio (WR) as the ratio of the number of images with unknown objects to the number of images with known objects. For setting T2, WR is gradually increased to construct three joint datasets: GraspNet-Test-{4, 5, 6}, containing {5120, 10240, 15360} images with WR = {1, 2, 3}. With these various test sets, we can comprehensively and unbiasedly study the OSIS performance of SemSeg.

As on COCO, we report AP_k and AOSE. But we use the Average Precision of unknown classes (AP_{unk}) as the performance metric of unknown objects.

The results on the GraspNet OSIS benchmark are shown in Tab. II and Tab. III. On all test sets in both test settings, SemSeg outperforms the baseline models across all metrics, especially for unknown objects, demonstrating the superiority of our approach. Notably, our model achieves almost the same level as Mask R-CNN on AP_k. Some qualitative results are illustrated in Fig. 4. SemSeg is able to segment known and unknown objects with high quality in cluster scenes, while OLN-Mask+PROSER misses many foreground objects and OpenDet+CA-Mask has a tendency to misclassify some unknown objects as one of the known classes.

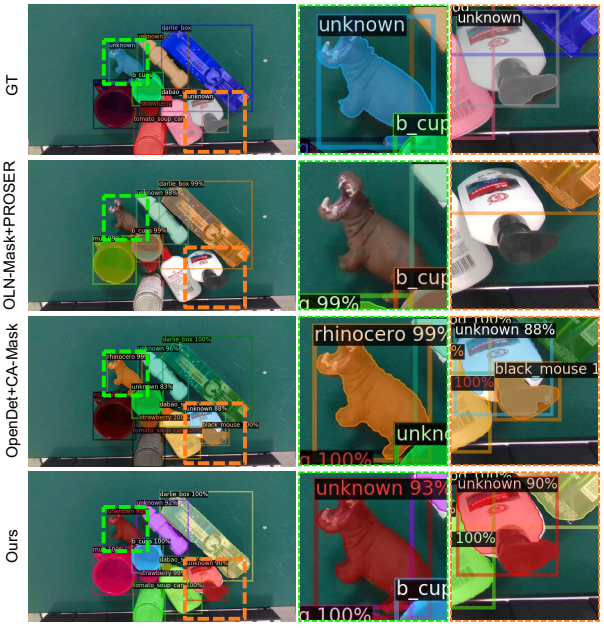


Fig. 4. Qualitative comparisons between SemSeg and other methods on the GraspNet OSIS benchmark. Known objects with confidence greater than 0.95 and unknown objects with confidence greater than 0.8 are visualized.

E. Evaluate Mask Quality

We conduct experiments on GraspNet-Test-3,6 to evaluate mask quality without interference from classification. In test, we use ground-truth bounding boxes and class labels to make the models predict masks only. Under this setting, Average Precision (AP) is only concerned with mask prediction, so AP_k and AP_{unk} can be used to quantify the mask quality of known and unknown objects respectively. The results are reported in Tab. IV showing the superiority of SemSeg. Fig. 5 shows some qualitative results which also demonstrate that our masks are generally of higher quality.

F. Analysis of Class Semantics Embedding Space

We use t-SNE to visualize the learned class semantics embedding space on GraspNet OSIS benchmark. In Fig. 6, the black dots represent embeddings of unknown objects and other colored dots represent embeddings of known objects from different known classes. The colored × denotes the center vector of each known class. The embeddings of known objects cluster around corresponding center vector and different known classes stand apart from each other. Most unknown object embeddings locate in the complementary region of known classes. And they do not cluster in one location, but have several clusters instead. So, it is rational to use the class semantics embedding space for open-set classification and class-semantics extraction.

TABLE IV
MASK QUALITY COMPARISON ON GRASPNET-TEST-3 AND GRASPNET-TEST-6. GIVEN GROUND-TRUTH BOUNDING BOXES AND CLASS LABELS, THE MODELS PREDICT ONLY MASKS DURING INFERENCE.

	GraspNet-Test-3		GraspNet-Test-6	
	$AP_k \uparrow$	$AP_{unk} \uparrow$	$AP_k \uparrow$	$AP_{unk} \uparrow$
OLN-Mask + PROSER	68.1	58.9	71.1	60.0
OpenDet + CA-Mask	68.2	57.2	71.1	58.8
SemSeg	**70.4**	**60.7**	**72.9**	**62.6**

TABLE V
ABLATION ON MAIN COMPONENTS (THE FIRST 4 ROWS) AND CLASS SEMANTICS UTILIZATION PATTERNS (THE LAST 2 ROWS) ON GRASPNET-TEST-6.

class semantics embedding space	class semantics modulation	AOSE ↓	$AP_k \uparrow$	$AP_{unk} \uparrow$
baseline		272834	57.3	0.0
	✓	287127	58.3	0.0
✓		52665	58.3	39.8
✓	✓	48885	59.1	43.0
✓	blend with input	53973	59.1	39.0

G. Ablation Studies

We perform ablation studies on GraspNet-Test-6 to analyze the effect of our main components and compare different class semantics utilization patterns. The results are shown in Tab. V. The baseline here is constructed by adding a softmax classifier to the class-agnostic instance segmentation model OLN-Mask.

As illustrated in the first four rows, adding class semantics modulation to the baseline improves the performance on known objects. Apart from the contribution to the segmentation of known objects, integrating class semantics embedding space enables the segmentation of unknown objects achieving AP_{unk} of 39.8 which is on par with or even better than the compared models in Tab. III (39.5 and 33.5). And it decreases AOSE alleviating the overconfidence problem. Our final model, the combination of the two modules, further enhances the performance, especially for unknown objects, which verifies the effectiveness of our proposed components.

The last two rows compare the two class semantics utilization patterns discussed in Section III-C: modulating network parameters or blending with network input. For the second technical route, we compute the cosine similarity between the proposal embedding and center vectors of all known classes in the embedding space. The similarities are combined into a similarity vector which is concatenated to the RoI feature map in the segmentation branch for mask prediction. While both methods perform similarly on known objects, blending with network input is inferior to modulating network parameters on unknown objects, even worse than the class-agnostic version (39.0 vs. 39.8). This is consistent with our analysis that blending class semantics with network input will result in severe difficulties with generalization.

H. Discussion on the Effect of Class-Specific Segmentation

Summarizing the results of all experiments, we observe an interesting phenomenon: no matter the comparison experiments on COCO and GraspNet or the ablation experiments, class-specific segmentation is much more beneficial for unknown objects than for known objects. We attribute this to the difference in generalization difficulty between known and unknown objects. Generalization from known objects in training to known objects in inference is straightforward, so the involvement of class semantics does not show a clear advantage. However, generalizing to unknown objects is nontrivial, making class semantics assistance imperative. This supports our theoretical analysis of the guidance role of class semantics in the segmentation process. Also, it validates the effectiveness of our design choices regarding the extraction and usage of class semantics, which are tailored for generalization to unknown objects.

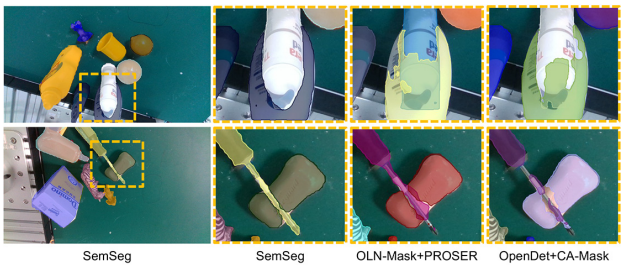


Fig. 5. Qualitative comparisons of mask quality using ground-truth bounding boxes and class labels.

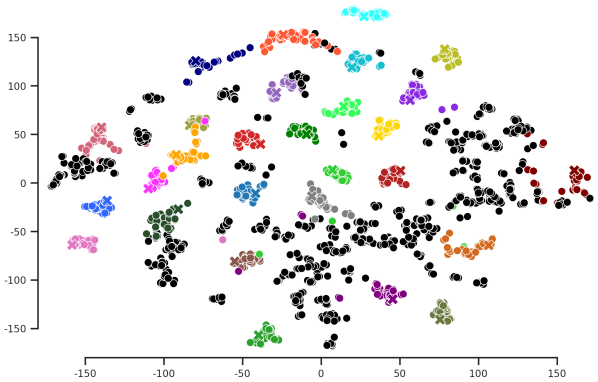


Fig. 6. t-SNE visualization of class semantics embedding space, including embeddings of unknown objects (black dots), embeddings of known objects from different known classes (other colored dots), and the center vector of each known class (colored ×).

V. CONCLUSION

We propose a novel method SemSeg to enable the class-specific mask prediction in open-set instance segmentation, which uses contrastive learning to construct a class semantics embedding space and utilizes the class semantics to modulate the segmentation branch. Class-specific OSIS allows optimizing the mask output layer for each class independently without competition between each other. Also, class semantics can guide and facilitate the segmentation task, which is essential for unknown objects with severe generalization challenges. Experiments on the COCO and GraspNet-1Billion datasets demonstrate the merits of our approach, especially on unknown objects.

REFERENCES

[1] J. A. Meacham, "Wisdom and the context of knowledge: Knowing that one doesn't know," *On the development of developmental psychology*, vol. 8, no. 111-134, p. 1, 1983.

[2] S. Engel, "Children's need to know: Curiosity in schools," *Harvard educational review*, vol. 81, no. 4, pp. 625–645, 2011.

[3] D. Miller, L. Nicholson, F. Dayoub, and N. Sünderhauf, "Dropout sampling for robust object detection in open-set conditions," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3243–3249.

[4] 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*, 2021, pp. 5830–5840.

[5] J. Han, Y. Ren, J. Ding, X. Pan, K. Yan, and G.-S. Xia, "Expanding low-density latent regions for open-set object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9591–9600.

[6] Z. Zhou, Y. Yang, Y. Wang, and R. Xiong, "Open-set object detection using classification-free object proposal and instance-level contrastive learning," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1691–1698, 2023.

[7] 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.

[8] K. Saito, P. Hu, T. Darrell, and K. Saenko, "Learning to detect every thing in an open world," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*. Springer, 2022, pp. 268–284.

[9] W. Wang, M. Feiszli, H. Wang, J. Malik, and D. Tran, "Open-world instance segmentation: Exploiting pseudo ground truth from learned pairwise affinity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4422–4432.

[10] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 7, pp. 1757–1772, 2012.

[11] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1563–1572.

[12] D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, "Learning placeholders for open-set recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4401–4410.

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

[14] A. Kirillov, Y. Wu, K. He, and R. Girshick, "Pointrend: Image segmentation as rendering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9799–9808.

[15] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang *et al.*, "Hybrid task cascade for instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4974–4983.

[16] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[17] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 444–11 453.

[18] J. Cen, P. Yun, J. Cai, M. Y. Wang, and M. Liu, "Deep metric learning for open world semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 333–15 342.

[19] R. Chan, K. Lis, S. Uhlemeyer, H. Blum, S. Honari, R. Siegwart, P. Fua, M. Salzmann, and M. Rottmann, "Segmentmeifyoucan: A benchmark for anomaly segmentation," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[20] J. Hong, W. Li, J. Han, J. Zheng, P. Fang, M. Harandi, and L. Petersson, "Goss: Towards generalized open-set semantic segmentation," *The Visual Computer*, pp. 1–14, 2023.

[21] J. Hwang, S. W. Oh, J.-Y. Lee, and B. Han, "Exemplar-based open-set panoptic segmentation network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1175–1184.

[22] H.-M. Xu, H. Chen, L. Liu, and Y. Yin, "Dual decision improves open-set panoptic segmentation," in *The 33rd British Machine Vision Conference (BMVC)*, vol. 2022, 2022, p. 3.

[23] D. Huynh, J. Kuen, Z. Lin, J. Gu, and E. Elhamifar, "Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7020–7031.

[24] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2955–2966.

[25] Z. Ding, J. Wang, and Z. Tu, "Open-vocabulary universal image segmentation with maskclip," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 8090–8102.

[26] R. Hu, P. Dollár, K. He, T. Darrell, and R. Girshick, "Learning to segment every thing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4233–4241.

[27] W. Kuo, A. Angelova, J. Malik, and T.-Y. Lin, "Shapemask: Learning to segment novel objects by refining shape priors," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9207–9216.

[28] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[29] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[30] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[31] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[32] T. Li, P. Cao, Y. Yuan, L. Fan, Y. Yang, R. S. Feris, P. Indyk, and D. Katabi, "Targeted supervised contrastive learning for long-tailed recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6918–6928.

[33] X. Zhao, R. Vemulapalli, P. A. Mansfield, B. Gong, B. Green, L. Shapira, and Y. Wu, "Contrastive learning for label efficient semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 623–10 633.

[34] 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*, 2019, pp. 9157–9166.

[35] Y. Wang, Z. Xu, H. Shen, B. Cheng, and L. Yang, "Centermask: single shot instance segmentation with point representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9313–9321.

[36] 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*, 2020, pp. 8573–8581.

[37] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 282–298.

[38] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "Solov2: Dynamic and fast instance segmentation," *Advances in Neural information processing systems*, vol. 33, pp. 17 721–17 732, 2020.

[39] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 213–229.

[40] 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, pp. 17 864–17 875, 2021.

[41] 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*, 2016, pp. 770–778.

[42] 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*, 2017, pp. 2117–2125.

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