SWORD ⚔: DEMYSTIFY THE SECRETS OF OPEN-WORLD INSTANCE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Current state-of-the-art instance recognition models have demonstrated strong ability in close-world environments while struggling in open-world scenarios, where the *novel* objects are not annotated in the pre-defined taxonomy during training. The challenge comes from that, in the unlabeled regions, *novel* objects and backdrop co-exist and are hard to differentiate. To demystify the secrets hidden in the mystery unannotated areas, we present a conceptually simple yet effective open-world instance recognition model, SWORD, answering the two critical questions: (1) How to discover the *novel* objects? We identify that the direct training of classification would make the features of *novel* objects degrade to the background. We demonstrate that a simple stop-gradient operation not only prevents feature degradation, but also allows the network to enjoy the merit of heuristic label assignment. (2) How to distinguish the objects from the backdrop? By maintaining a universal object queue, we obtain the object center for performing contrastive learning, in order to enlarge the distinction between objects and background. While the previous works only focus on pursuing recall and neglect precision, we show the prominence of SWORD by giving consideration to both criteria and achieving state-of-the-art performance in various open-world cross-category and cross-dataset generalizations. In particular, on VOC to non-VOC setup, our method sets a new state-of-the-art of 39.6% on $\text{AR}^{\text{b}}_{100}$. For COCO to UVO generalization, SWORD significantly outperforms the previous best open-world model by 6.0% on AP^{b} and 9.0% on $\text{AR}^{\text{b}}_{100}$, respectively.

1 INTRODUCTION

Instance recognition (*i.e.*, object detection and instance segmentation) is one of the fundamental tasks in computer vision and the deep learning-based methods have progressed drastically with advanced techniques. However, the modern instance recognition methods (Ren et al., 2015; He et al., 2017; Tian et al., 2019; 2020; Wang et al., 2020; Cheng et al., 2021; 2022) are ideally based on the close-world assumption, *i.e.*, they are designed to detect the objects in the same pre-defined fixed taxonomy in both training and inference phases. Despite the rapidly growing size of object categories in today's large-scale datasets (Shao et al., 2019; Kuznetsova et al., 2020), it is hardly possible to comprehensively cover all the objects in the real world. While in practice, there exist many scenarios where the instance recognition models would encounter *novel* and *unknown* objects at inference time, *e.g.*, the autonomous driving and robotic manipulation. Towards building more generalized artificial intelligent systems (Goertzel, 2014), it has an imperative need to develop models that possess the open-world instance recognition ability, *i.e.*, the network could localize any objects in the images while only trained on the partial annotations of limited object categories.

In the open-world scenario, the common drawback of current close-world models is that they regard all the unlabeled regions as *background* during training, and thus the classification head would assign low scores to those *unknown* objects without annotations. To mitigate the issue, Kim et al. (2022) propose the classification-free Object Localization Network (OLN), a variant of standard Mask-RCNN (He et al., 2017). OLN removes the classification head and learns to predict the scores of object proposals with the localization quality head (Tian et al., 2019). In this manner, the *unknown* objects would not be suppressed since only the positive samples overlapped with ground-truth are trained to estimate the localization quality scores, making the network able to discover the *novel*

objects. Despite the promising results on average recall (AR), OLN fails to perceive the backdrop and produces numerous false positive predictions, resulting in fairly low average precision (AP).

Considering that open-world instance recognition is fundamentally a problem of learning the generalization from *base* objects to *novel* objects, it has the essential requirement: the network needs to learn the common characteristics of objects and backdrop as well as distinguish them. Although OLN (Kim et al., 2022) gives the insight that localization quality is a generalizable objectness score, it is lack of discriminative ability due to the absence of negative samples during training. We argue that **the secrets of open-world instance recognition still hide in the mystery unlabeled regions where the *novel* objects and backdrop usually co-exist and are hard to differentiate**. And this causes two great challenges: **(1)** How to discover the *novel* objects? **(2)** How to distinguish the objects from the backdrop? In this work, we propose SWORD, unsealing the secrets of open-world instance recognition and cutting off the obstacles with two swords.

For the first sword, we propose to attach a `stop-grad` operation before the classification head. We identify that the intrinsic reason why the close-world models perform poorly in the open-world setting is that the features of *novel* objects degrade to background due to the foreground/background learning of classification head. With the simple yet decisive operation, we make the *novel* objects appear in the feature maps again. Moreover, the preserve of classification scores not only helps remedy the weak distinguishability of IoU scores, but also allows the network to enjoy the merit of heuristic label assignment.

For the second sword, we design a novel contrastive learning framework for learning the discriminative representations between objects and backdrop. The core idea is to ensure similar representations among objects while enlarging the distinction between the objects and backdrop in the feature space. Specifically, we maintain a universal object queue to store the annotated object embeddings. The pooling feature of the queue, *i.e.*, object center, captures the common characteristics of objects and plays as the role of *query* in contrastive learning. The previous works (Kim et al., 2022; Konan et al., 2022) select the positive and negative samples by IoU threshold, which would introduce many false positives. Different from theirs, we formulate the sample selection problem as the optimization problem for optimal transport (Villani, 2009; Peyré et al., 2019). And we only select those hard examples based on the matching cost as negative pairs to improve the quality of embeddings. Contrastive learning is the key to reducing false positives and greatly improves precision.

By combining the two swords, our SWORD has following appealing advantages: **(1)** It is a unified and simple framework, totally getting rid of the hand-crafted components (*e.g.*, anchor design and IoU threshold setting). **(2)** By discriminating the objects and backdrop, it not only reveals the strong ability in recalling *novel* objects, but also achieves high precision. **(3)** The open-world knowledge of SWORD could be easily transferred to the standard instance recognition models.

The contributions of this paper are as follows: **(1)** We point out that the key to the success of open-world instance recognition lies in preventing the disappearance of *novel* objects in feature maps and learning the discrimination between objects and backdrop. **(2)** We propose a simple yet effective open-world instance recognition framework, SWORD, which shows excellent ability in recalling *novel* objects without inducing many false positives. **(3)** Extensive experiments demonstrate that our models achieve the state-of-the-art performance in various open-world cross-category and cross-dataset settings on several benchmarks including COCO (Lin et al., 2014), LVIS (Gupta et al., 2019), UVO (Wang et al., 2021a) and Objects365 (Shao et al., 2019).

2 Related Work

Open-world Instance Recognition. Towards building more practical applications in the real world, the open-world-related problems (Bendale & Boult, 2015; Cen et al., 2021; Han et al., 2019; Cen et al., 2021; Qi et al., 2021) have raised great attention recently. Kim et al. (2022) firstly establish the protocol of open-world proposal (OWP) problem, which is also named open-world instance recognition. Literally, the model needs to produce class-agnostic box or mask proposals to localize all the objects in the images, while only annotations of partial object categories are available. There are several works attempting to solve the problem from various aspects, *e.g.*, OLN (Kim et al., 2022), LDET (Saito et al., 2021) and GGN (Wang et al., 2022a). Please see Appendix A for a comprehensive review of these works.

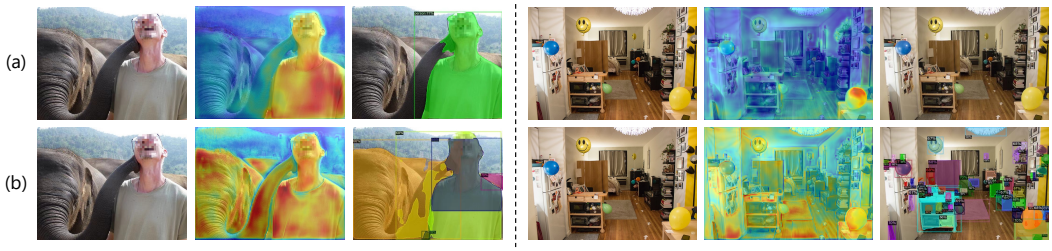


Figure 1: **The visualization results of (a) close-world Deformable-DETR and (b) open-world OLN**. For each example, we show the input image, feature map and predicted result from left to right. Note that the 'elephant', 'refrigerator' and 'bed' are not annotated in the training set. Deformable-DETR can not discover the *novel* objects while OLN produces many false positives.

Contrastive Learning. Contrastive learning has been dominant in both self-supervised (He et al., 2020; Chen et al., 2020; Grill et al., 2020; Chen & He, 2021; Xie et al., 2021; Wang et al., 2021c; Khosla et al., 2020; Bai et al., 2022) and supervised (Khosla et al., 2020; Han et al., 2020; Wu et al., 2022; Pang et al., 2021) representation learning. The core idea lies in that the positive samples are attracted while the negative samples are pulled away in the feature space to learn the discriminative representations. MOCO (He et al., 2020) maintains a memory queue to store a large number of negative pairs and enables the momentum update of the memory encoder to guarantee the queue feature consistency. SimSiam (Chen & He, 2021) develops the extremely simple siamese network without any negative sample, and points out a `stop-grad` operation plays an essential role in preventing mode collapse. In this work, we absorb the ideas from contrastive learning to learn the distinct representations of objects and backdrop.

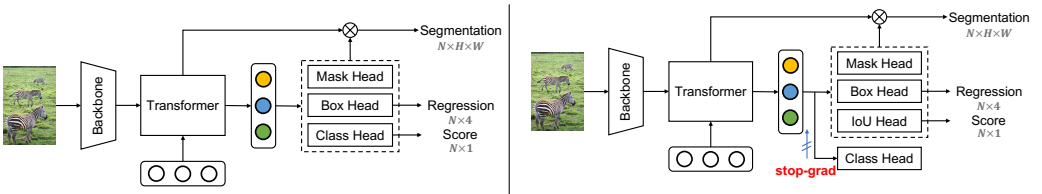
3 Method

3.1 Problem Definition

An open-world instance segmentation model can not only segment all the previously *known* objects, but also recognize the *unknown* instances during inference. Formally, we formulate the open-world instance segmentation problem in the following canonical form. Given an instance segmentation dataset $\boldsymbol{D}_1$, we have the partial object annotations on the *base* category set $\mathcal{C}_{in} = \{c_1, c_2, ..., c_k\}$. Notably, there are also a large number of *unknown* objects co-appearing in the data while remaining unlabeled. The model are trained to provide a set of class-agnostic proposals $\mathcal{P} = \{s_i, b_i, m_i\}_{i=1}^{p}$ to localize *all* objects in the image, where $s_i \in \mathbb{R}$ indicates the proposal confidence, $b_i \in \mathbb{R}^4$ denotes the bounding box coordinates and $m_i \in \mathbb{R}^{H \times W}$ is the segmentation mask. During inference, the model could be evaluated either on the same dataset $\boldsymbol{D}_1$ or another dataset $\boldsymbol{D}_2$. The important generalization ability of the model is revealed by recalling the *novel* objects which are in the orthogonal category set $\mathcal{C}_{out} = \{c_{k+1}, c_{k+2}, ..., c_n\}$.

3.2 From Close-world to Open-world Instance Recognition

Discussion. To dig out the secrets that restrict the development of open-world instance recognition, we visualize the predicted results of a close-world model (Deformable-DETR (Zhu et al., 2020)) and an open-world model (OLN (Kim et al., 2022)) in Figure 1. On one hand, we notice that the features of *novel* objects are degraded to the background for the close-world model, which we term as *feature degradation*. As shown in the left example of Figure 1(a), the elephants are not annotated in the training set and their features can hardly be distinguished from their surroundings. This is because the classification head is trained to identify the foreground and background. Since the *novel* objects are unannotated, they are also treated as the background during training. On the other hand, OLN replaces the classification head with the localization quality head to estimate the scores of proposals. In this manner, despite the *novel* objects would not be suppressed, the network loses the perception of the backdrop and lacks the discriminative ability since only the positive samples are trained. As shown in Figure 1(b), it produces many false positive predictions, *e.g.*, parts of the man's body in the first example and overlapped tables in the second example. To conclude, there are two critical issues for the open-world instance recognition: **preventing the *feature degradation* of *novel* objects** and **learning the discrimination between objects and backdrop**. In this work, we propose SWORD, aiming to cut off the two obstacles with two swords.



(a) Close-world Transformer (b) Open-world Transformer

Figure 2: Our networks are based on the (a) Deformable-DETR. The instance segmentation is achieved by adding the dynamic mask head (Tian et al., 2020). (b) By using the scores of the IoU head and attaching the `stop-grad` operation before the classification head, the close-world Transformer is transformed into an open-world Transformer.

Open-world Transformer. To this end, we first develop a simple open-world instance recognition model. Our networks are built upon the recent Deformable-DETR (Zhu et al., 2020) due to its simple architecture. First, we add the mask head on top of the Transformer to generate the instance mask by performance dynamic convolution (Tian et al., 2020; Cheng et al., 2022), which is shown in Figure 2a. Then, the model is transformed into an simple open-world model (Figure 2b) for learning the open-world class-agnostic mask proposals by making the following modifications.

- IoU Score as Proposal Confidence. As pointed out by Kim et al. (2022), the localization quality is a better confidence cue than the classification score in the open-world setting. Inspired by this philosophy, we add the extra two IoU heads on top of the Transformer decoder to predict the box IoU score c_b and mask IoU score c_m, respectively. During inference, we use the geometric mean of IoU scores, *i.e.*, $s = \sqrt{c_b \cdot c_m}$, as the proposal confidence. In Table 1, we show the results of recalling *novel* objects on VOC to non-VOC setting. By comparing the first and third rows in Table 1, we could observe that AR_{100} is greatly boosted by using the IoU score as proposal confidence.

- Stop-gradient Operation. Different from OLN (Kim et al., 2022) that discards the classification head, we preserve the classification head and show the first sword to prevent the *feature degradation*. Here, we propose a simple yet effective solution: attaching a `stop-grad` operation before the classification head. On one hand, it prevents the gradient passing from the classification head to the network parameters so that the unlabeled regions would not be suppressed. On the other hand, it can be seamlessly applied on the advanced DETR-like detectors (Carion et al., 2020; Meng et al., 2021; Liu et al., 2022; Li et al., 2022) and enables the heuristic label assignment (Sun et al., 2021) which considers the classification cost. The effectiveness of this design is validated in Table 1. Even though the classification head is still trained to recognize the *base* objects as foreground while others as background, $\text{AR}^{\text{b}}_{100}$ on *novel* objects gets 5.8% gain.

Table 1: The results of using different confidence scores and w/o stop-grad operation on VOC to non-VOC setup.

confidence	stop-grad	$\text{AR}^{\text{b}}_{100}$	$\text{AR}^{\text{m}}_{100}$
Class	✗	18.2	13.5
	✓	24.0 (+5.8)	18.8 (+5.3)
IoU	✗	23.6	18.8
	✓	29.3 (+5.7)	24.7 (+5.9)

3.3 SWORD: Learning to Discriminate Objects and Backdrop

In this subsection, we propose a contrastive learning framework for learning the distinct representations of objects and backdrop, which is the second sword towards open-world instance recognition.

Contrastive Learning between Objects and Backdrop. Analogous to OLN (Kim et al., 2022), open-world Transformer is lack of discrimination for separating the foreground and background. To solve the challenge, we introduce contrastive learning (He et al., 2020; Grill et al., 2020; Chen & He, 2021) to learn the more discriminative features between objects and backdrop.

As shown in Figure 3, we further add the contrastive head on top of the open-world Transformer decoder to learn the query embeddings. Moreover, we maintain a universal object queue to store the object embeddings, which come from those queries best matching the ground-truth objects. Notably, these embeddings are encoded by a slowly progressing contrastive head, *i.e.*, the parameters are updated by the exponential moving average (EMA) method:

$$\theta_c^{'} \leftarrow \alpha\theta_c^{'} + (1-\alpha)\theta_c \tag{1}$$

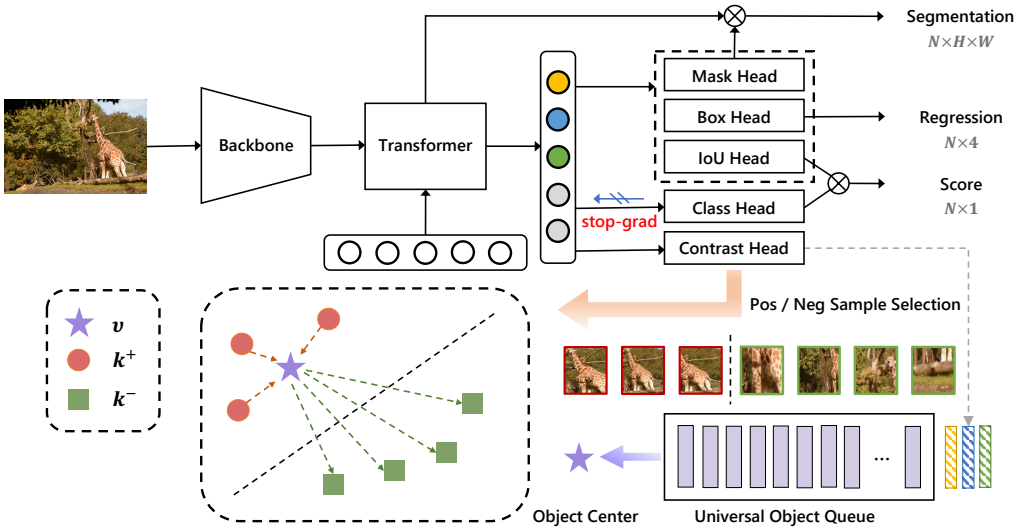


Figure 3: **The overall framework of SWORD**. The network is built upon the open-world Transformer introduced in Sec. 3.2. A contrastive head is further added on top of the decoder to predict the query embeddings. We maintain a universal object queue to store the embeddings of annotated objects. The pooled feature, *i.e.*, object center, captures the common object characteristics and plays the role of query in contrastive learning. The positive and negative pairs are selected from query embeddings dynamically by the optimal transport process for each ground-truth object.

where θ_c denotes the parameters of the regularly updated contrastive head. We pool all the features in the universal object queue to get the object center v, which plays as the query in contrastive learning. Intuitively, the object center is the common object representation and has two appealing advantages: (1) *discriminativeness*: the features in the object queue are from those annotated objects, which could be easily distinguished from the unlabeled regions. (2) *consistency*: the object center stays stable in the feature space thanks to the EMA update. Suppose now we have obtained the positive (objects) embeddings $\mathcal{K}^+$ and negative (background) embeddings $\mathcal{K}^-$, we expect that the positive sample features should be close to the object center while the negative ones should be pulled away. The contrastive loss is defined as follows:

$$\mathcal{L}_{con} = -\log \frac{\sum_{k^+ \in \mathcal{K}^+} \exp(v \cdot k^+)}{\sum_{k^+ \in \mathcal{K}^+} \exp(v \cdot k^+) + \sum_{k^- \in \mathcal{K}^-} \exp(v \cdot k^-)} \tag{2}$$

Positive and Negative Sample Selection. We formulate the sample selection problem as the optimal transport problem (Ge et al., 2021; Villani, 2009; Peyré et al., 2019) to automatically select the positive and negative samples for contrastive learning. Specifically, we take the classification results into consideration and compute the cost between the predictions and ground-truths:

$$C = \lambda_{cls} \cdot C_{cls} + \lambda_{box} \cdot C_{box} \tag{3}$$

where, C_{cls} is Focal loss (Lin et al., 2017), and C_{box} is a combination of the $\mathcal{L}_1$ loss and generalized IoU loss (Rezatofighi et al., 2019). We demonstrate that adding the classification cost would help the network choose more discriminative samples. Ideally, the predictions with the least cost are those objects close to the ground-truths. To improve the quality of learned embeddings, we first dynamically choose k_1 and k_2 predictions with the least cost, where $k_2 > k_1$. Then the k_1 predictions are positive samples, and the left $k_2 - k_1$ predictions are hard negatives.

Training. The label assignment for the network optimization also relies on the matching cost as Eq. (3). The k_1 predictions are also considered as the positive samples for the network training and other predictions are all negative samples. The overall loss function for training is:

$$\mathcal{L} = \lambda_{cls} \cdot \mathcal{L}_{cls} + \lambda_{box} \cdot \mathcal{L}_{box} + \lambda_{mask} \cdot \mathcal{L}_{mask} + \lambda_{iou} \cdot \mathcal{L}_{iou} + \lambda_{con} \cdot \mathcal{L}_{con} \tag{4}$$

where, $\mathcal{L}_{cls}$ is computed on all predicted samples and $\mathcal{L}_{box}, \mathcal{L}_{mask}, \mathcal{L}_{iou}$ are applied on the positive samples. The mask-related loss is a combination of the DICE loss (Milletari et al., 2016) and Focal loss (Lin et al., 2017). The IoU scores are supervised by the binary cross entropy loss $\mathcal{L}_{iou}$.

Inference. As the IoU heads are only trained with positive samples and thus they will assign high scores for all the proposals, which would cause a large number of false positive (FP) predictions. So we use the multiplication results of classification scores and IoU scores as final scores, *i.e.*, $s = \sqrt[3]{c_c \cdot c_b \cdot c_m}$. And NMS post-process is applied to remove the redundant predictions.

3.4 SWORD*: Teacher-student Knowledge Transfer

The close-world models fail to localize the *unknown* objects as they would consider all the unannotated regions as background. We demonstrate that by receiving the knowledge from SWORD, the close-world models could also learn to detect the *novel* objects. Therefore, we further develop teacher-student learning to improve the generalization ability of close-world models. As shown in Figure 5, the pretrained SWORD is employed to generate pseudo labels on the partially-annotated training images. Then we merge the original annotations with the top-k predicted proposals which have low overlap with base objects to form the augmented annotations. Finally, the standard Deformable-DETR (Zhu et al., 2020) is trained under the supervision of updated annotations, which we term as SWORD*. Surprisingly, teacher-student learning can effectively boost the close-world model's ability to localize *novel* objects, as presented in Sec. 4. In the training process, the teacher model is kept frozen and we empirically find that using the IoU scores for proposals of the teacher model has a better learning effect. Please see more details in Appendix C.

4 Experiments

In this section, we first thoroughly evaluate the performance of SWORD in two challenging settings, including the cross-category and cross-dataset generalizations. Then we conduct extensive ablation studies to discuss the key designs and analyze the crucial issues in Sec. 4.4.

4.1 Experiment Settings

Datasets. Our experiments are conducted on COCO (Lin et al., 2014), LVIS (Gupta et al., 2019), UVO (Wang et al., 2021a) and Objects365 (Shao et al., 2019) datasets. **COCO** is the widely used instance segmentation benchmark with 80 categories. **LVIS** shares the same images with COCO while having a more complete label system. It has a large taxonomy of 1203 categories in a long-tailed distribution. **UVO** originates from the Kinetics400 (Carreira & Zisserman, 2017) dataset and all the instance masks are exhaustively annotated. **Objects365** is a large-scale object detection dataset with 2 million images and 365 categories where all the COCO 80 categories are included.

In the experiments, we consider two challenging open-world settings: **(1) Cross-category evaluation**. It means that the models are trained and evaluated on the same dataset while only partial annotations of *base* objects are available during training. On COCO benchmark, we follow the common practice (Kim et al., 2022; Saito et al., 2021) to split the annotations into two non-overlapping class sets, where the Pascal VOC (Everingham et al., 2010) 20 classes are adopted as the *base* set and other 60 non-VOC classes are *novel* set. On LVIS dataset, we train the models on the COCO 80 categories and use the rest non-COCO categories for evaluation. The results are reported on the *novel* objects. **(2) Cross-dataset evaluation**. To test the model's generalization ability to new environments, we use COCO as the training source and evaluate the models on new datasets, *i.e.*, UVO [1] and Objects365[2]. In this setting, we not only include the results on *novel* categories, but also on all categories to test the model's domain generalization in the wild.

Evaluation Metrics. All the models are evaluated in a class-agnostic way, and we use the average recall (AR) and mean average precision (mAP) over multiple IoU thresholds $[0.5 : 0.95]$ as the standard metrics to measure the performance of models.

[1]The downsampled dense split of v1.0 contains two classes: "objects" for COCO categories and "other" for non-COCO categories. The NOVEL metrics are measured on the "other" categories. The previously released v0.5 does not distinguish the object categories and all the objects are annotated as "objects". We report the results of ALL metrics based on this version.

[2]Objects365 only has box annotations, so we report the results regarding the box metrics.

Table 2: **Comparison of state-of-the-art performance on VOC to non-VOC setting.**

Method	AP^b	AR^b_{10}	AR^b_{100}	AP^m	AR^m_{10}	AR^m_{100}
Mask-RCNN	1.6	10.2	23.5	0.9	7.9	17.7
OLN	3.7	18.0	33.5	-	16.9	-
LDET	5.0	18.2	30.8	4.3	16.3	27.4
GGN	5.8	17.3	31.6	4.9	16.1	28.7
GGN + OLN	3.4	17.1	37.2	3.2	16.4	33.7
Deformable-DETR	1.1	7.6	18.2	0.7	5.8	13.5
SWORD	5.8	17.8	35.3	4.8	15.7	30.2
SWORD*	**6.3**	**21.7**	**39.6**	**5.6**	**20.0**	**34.5**

Table 3: **Comparison of state-of-the-art performance on COCO to LVIS setting.** We use the LVIS annotations on COCO-category objects for training.

Method	AR^b_{10}	AR^b_{100}	AR^m_{10}	AR^m_{100}
Mask-RCNN	6.1	19.4	5.6	17.2
GGN	-	-	-	20.4
Deformable-DETR	6.3	19.4	5.5	16.4
SWORD	8.8	23.5	8.0	20.4
SWORD*	**9.7**	**26.5**	**9.0**	**22.7**

Implementation Details. We use ResNet-50 (He et al., 2016) as the backbone by default and follow the same model setting as Deformable-DETR, *i.e.*, the network has 6 encoders and 6 decoders with the hidden dimension of 256. The size of the universal object queue is set as 4096 and the EMA rate α is 0.999. Please see more implementation details in Sec. B.

4.2 Cross-category Evaluation

VOC to non-VOC. The cross-category generalization in VOC to non-VOC setting is a challenging problem because only a small-sized taxonomy (20 classes) are available in the training set. We compare our methods with other state-of-the-art methods in Table 2. It could seen that SWORD achieves the significant 17.1% AR^b_{100} gain and 16.7% AR^m_{100} gain compared with the Deformable-DETR baseline. And SWORD outperforms the all previous single model, *e.g.*, the improvement is +1.8% on AR^b_{100} and +1.5% on AR^m_{100}. Interestingly, by incorporating teacher-student learning, the performance can be further greatly boosted. SWORD* achieves state-of-the-art performance in all metrics. It surpassing the previous best GGN + OLN by 2.4% AR^b_{100} and 0.8% AR^m_{100}.

COCO to LVIS. The results of COCO to LVIS setting are shown in Table 3. SWORD has obvious performance improvement over the Deformable-DETR baseline (23.5% v.s. 19.4% on AR^b_{100}, 20.4% v.s. 16.4% on AR^m_{100}). Table 3 also demonstrates that teacher-student learning benefits the close-world model to localize *novel* objects. SWORD* creates a new state-of-the-art result and surpasses Deformable-DETR up to 7.1% and 6.3% in terms of AR^b_{100} and AR^m_{100}, respectively.

4.3 Cross-dataset Evaluation

COCO to UVO. The COCO to UVO setting has the obvious domain shift since the UVO dataset originates from the video dataset Kinetics400 (Carreira & Zisserman, 2017), which makes the setup suitable for evaluating the model's generalization ability in the wild. We thoroughly compare the results on *novel* objects and *all* objects in Table 4. The baseline Deformable-DETR demonstrates strong performance in this setting and outperforms the previous methods by a large margin. Our SWORD further improves the performance over the strong baseline for all the metrics. For instance, +2.9% AP^b and +1.2% AR^m_{100} for *all* objects. The performance advantage is more clear for the *novel* objects, *e.g.*, SWORD shows 3.3% AP^b and 2.7% AR^m_{100} gain compared with Deformable-DETR.

Another interesting observation is that the teacher-student learning still proves to be effective for the model to localize *novel* objects and improves the proposal recall, nevertheless, it would decrease the average precision (AP) on *all* objects. By comparing the third-to-last row and the last row of Table 4, we could see that AP^b drops from 29.1% to 28.4%. The reason attributes to that the model is trained to produce more high-score proposals, and some of them are false positive predictions.

COCO to Objects365. We further conduct another cross-dataset evaluation on COCO to Objects365 and the results are listed in Table 5. The Mask-RCNN-based method LDET[3] improves Mask-RCNN in terms of AR while decreasing AP. SWORD obviously outperforms Deformable-DETR baseline for all metrics, which proves the superiority of our method. The last row in Table 5 illustrates that the pseudo-labeling training can significantly improve the recall ability of the network. Considering that SWORD* has exactly the same architecture as Deformable-DETR, it is impressive that AR_{100} on *novel* objects enjoys the 4.9% gain by this process.

[3]We report the results of LDET using the same class-agnostic evaluation for fair comparison, whereas results in the original paper are based on the class-wise evaluation.

Table 4: Comparison of state-of-the-art performance on COCO to UVO setting.

Method	Novel						All					
	AP^b	AR^b_{10}	AR^b_{100}	AP^m	AR^m_{10}	AR^m_{100}	AP^b	AR^b_{10}	AR^b_{100}	AP^m	AR^m_{10}	AR^m_{100}
Mask-RCNN	11.8	16.4	30.4	7.0	13.8	25.5	25.7	30.2	43.8	20.7	25.7	36.7
LDET	12.9	19.0	35.9	8.2	15.9	30.5	26.0	30.9	47.0	22.1	27.3	40.7
GGN	-	-	-	-	-	-	24.0	29.8	52.2	20.3	-	43.4
Deformable-DETR	14.2	20.0	45.8	9.0	16.7	37.9	29.1	35.0	60.7	24.7	30.1	50.3
SWORD	**17.5**	22.2	48.1	**12.8**	19.4	40.6	**32.0**	**36.5**	61.2	**28.0**	**32.4**	51.5
SWORD*	16.7	**22.6**	**49.8**	12.8	**20.8**	**42.4**	28.4	35.2	**62.5**	25.7	32.3	**52.7**

Table 5: Comparison of state-of-the-art performance on COCO to Objects365 setting.

Method	Novel						All					
	AP^b	AR^b_{10}	AR^b_{100}	AR^b_s	AR^b_m	AR^b_l	AP^b	AR^b_{10}	AR^b_{100}	AR^b_s	AR^b_m	AR^b_l
Mask-RCNN	13.0	19.3	32.8	18.2	36.4	43.5	25.1	23.9	40.3	22.7	42.8	53.4
LDET	12.8	20.0	36.8	20.7	40.5	48.9	22.5	22.7	41.4	22.9	44.3	54.9
Deformable-DETR	12.9	19.0	40.1	22.8	43.4	54.1	27.3	25.3	48.7	27.5	50.9	65.6
SWORD	**16.6**	22.8	43.9	**25.0**	48.6	57.6	**29.7**	**27.3**	50.8	**28.6**	54.0	67.2
SWORD*	16.0	**23.6**	**45.0**	23.9	**49.3**	**61.5**	27.8	26.9	**51.2**	27.2	**54.3**	**69.5**

4.4 Ablation Study

In this subsection, we conduct extensive ablation studies to analyze the crucial composing of our method. The experimental results are based on the COCO to UVO setting and we use the ResNet50 as backbone otherwise specified. The results are reported in terms of box metrics.

Analysis of Key Designs. Table 6 summarizes the results to study the key designs of our method. We also report the results of Deformable-DETR (Zhu et al., 2020) for comparison in the first row. We notice that Deformable-DETR shows great performance on *all* objects. This is because the training source, *i.e.*, COCO dataset, has provided the exhaustive annotations of diverse object categories so that Deformable-DETR could perform considerable well on these in-taxonomy objects.

Towards building an open-world instance segmentation model, we start from the open-world Transformer and gradually add the crucial components. First, by introduce the `stop-grad` operation, the performance is slightly improved. This operation is more critical for the cross-category generalization where the *novel* objects would be suppressed, as illustrated in Table 1. Second, the advanced one-to-many label assignment is adopted, which has been demonstrated the effectiveness in many previous works (Wang et al., 2022b; Jia et al., 2022; Hong et al., 2022). At this point, the network has achieved the competitive even better performance on *novel* objects compared with Deformable-DETR. However, the results on *all* objects are still far from satisfaction. To reduce the false positive predictions, we multiply the IoU scores with classification scores as proposal confidences during inference. And we observe the AP^b for *all* objects is obviously improved from 22.1% to 25.1%. Lastly, the proposed contrastive learning strategy fundamentally enforces the network to learn the discrimination between objects and backdrop. We are surprised to see that AP^b for *all* objects further obtains 6.9% gain and all the metrics surpass open-world Transformer by a large margin.

Classification Cost for Sample Selection in Contrastive Learning. The selection of positives and negatives is crucial for contrastive learning. To evaluate the effect of classification cost in this process, we set $C_{cls} = 0$ in Eq. (3) for the ablation. From Table 7, we observe that performance drops drastically without classification cost. Such phenomenon stands with the view that classification score is important and it has two potential reasons. First, the classification cost ensures the network's behavior consistency of label assignment for both contrastive learning and network training. Second, the localization cost alone will introduce those predictions closest to the ground-truths as positive samples, while classification cost helps choose more discriminative samples.

Do Stronger Backbones Benefit in Open-world? There exists the consensus that stronger backbones (He et al., 2016; Dosovitskiy et al., 2020; Wang et al., 2021b; Liu et al., 2021b; Tolstikhin et al., 2021; Chen et al., 2021) could greatly increase the performance under the fully-supervised setup. Of particular interest, we examine with ResNet (He et al., 2016) and Swin-Transformer (Liu

Table 6: **Ablation on the key designs of our method**. We start from the open-world Transformer as baseline and gradually add the key components. The last row is the full model of SWORD.

Variants	Novel			All		
	AP^b	AR^b_{10}	AR^b_{100}	AP^b	AR^b_{10}	AR^b_{100}
Deformable-DETR	14.2	20.0	45.8	29.1	35.0	60.7
Open-world Transformer	11.3	17.8	45.8	17.5	28.0	58.1
+ stop-grad	12.0 (+0.7)	18.5 (+0.7)	46.0 (+0.2)	18.5 (+1.0)	28.5 (+0.5)	58.4 (+0.3)
+ one-to-many assignment	14.1 (+2.1)	20.3 (+1.8)	48.6 (+2.6)	22.1 (+3.6)	32.0 (+3.5)	61.0 (+2.6)
+ class score	15.1 (+1.0)	21.6 (+1.3)	49.0 (+0.4)	25.1 (+3.0)	34.2 (+2.2)	61.2 (+0.2)
+ contrastive learning	17.5 (+2.4)	22.2 (+0.6)	48.1 (−0.9)	32.0 (+6.9)	36.5 (+2.3)	61.2 (+0.0)

Table 7: Ablation on the classification cost for sample selection in contrastive learning.

class cost	Novel			All		
	AP^b	AR^b_{10}	AR^b_{100}	AP^b	AR^b_{10}	AR^b_{100}
✗	13.7	20.1	48.3	19.8	29.5	60.2
✓	17.5	22.2	48.1	32.0	36.5	61.2

Table 8: Ablation on the backbones.

backbone	Novel			All		
	AP^b	AR^b_{10}	AR^b_{100}	AP^b	AR^b_{10}	AR^b_{100}
R50	17.5	22.2	48.1	32.0	36.5	61.2
R101	17.7	22.7	48.7	33.9	37.6	62.1
Swin-T	17.2	22.5	48.1	33.7	37.7	61.7
Swin-L	18.7	23.3	47.8	38.0	40.9	62.6

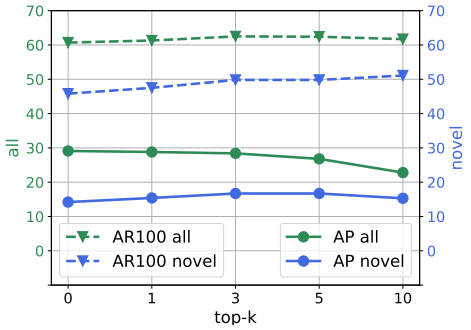


Figure 4: The effect of pseudo label number in teacher-student learning.

et al., 2021b) to study the effect of using strong backbones in open-world scenario. Table 8 illustrates that model consistently performs better with increasing the size of backbones. Interestingly, we also observe that out-of-taxonomy objects gets less benefit from stronger backbone than in-taxonomy objects in the open-world environment. For example, the model enjoys the significant 4.3% AP^b gain for *all* objects while the advance is marginal for *unknown* objects (+1.5% AP^b) by switching the backbone from Swin-Tiny to Swin-Large.

How Many Pseudo Labels are Needed in Teacher-student Learning? The usage of pseudo labels in teacher-student learning helps discover the *un-annotated* objects, on the other hand, it also introduces noisy supervision signals. To evaluate the relationship between the model behavior and pseudo labels, we vary the number of top-k for selecting pseudo labels and plot the results in Figure 4. Here, we have two critical findings: (1) The choose of top-k is a trade-off between *novel* objects and *base* objects. We can observe that AR_{100} for *novel* objects achieves the best result when $k = 5$ while not for the *all* objects. Intuitively, the introduction of pseudo labels will enforce the model to produce more high-scoring proposals for localizing the *novel* objects, which results in fewer proposals for the *base* objects inversely. (2) More pseudo labels benefit AR while hurting AP. As shown in Figure 4, AR_{100} keeps improving with the increase of k. On the contrary, AP for *all* objects consistently degrades. And AP for *novel* objects also starts decreasing when k is set as a large value (*e.g.*, $k = 10$). This is reasonable because more pseudo labels will induce many false positive predictions, which is harmful to network training. Therefore, the value of top-k should be carefully chosen to achieve the optimal balance for all the criteria.

5 Conclusion

In this work, we present a novel framework, SWORD, for open-world instance recognition. Specifically, we identify that a `stop-grad` operation is the key to preventing the feature degradation of *novel* objects and propose a contrastive learning strategy to enlarge the distinction between objects and backdrop. We further develop SWORD*, illustrating that a standard close-world model could also perform favorably well in the open-world setups by receiving the knowledge from SWORD. Experimental results demonstrate that the proposed models achieve the state-of-the-art performance on various cross-category (*e.g.*, VOC to non-VOC, COCO to LVIS) and cross-dataset (*e.g.*, COCO to UVO, COCO to Objects365) generalization.

REFERENCES

Pablo Arbelaez. Boundary extraction in natural images using ultrametric contour maps. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pp. 182–182. IEEE, 2006.

Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010.

Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 328–335, 2014.

Yutong Bai, Xinlei Chen, Alexander Kirillov, Alan Yuille, and Alexander C Berg. Point-level region contrast for object detection pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16061–16070, 2022.

Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1893–1902, 2015.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. Deep metric learning for open world semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15333–15342, 2021.

Shoufa Chen, Enze Xie, Chongjian Ge, Ding Liang, and Ping Luo. Cyclemlp: A mlp-like architecture for dense prediction. *arXiv preprint arXiv:2107.10224*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1290–1299, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 303–312, 2021.

Ben Goertzel. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1, 2014.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

Agrim Gupta, Piotr Dollar, and Ross 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.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8401–8409, 2019.

Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Qinghang Hong, Fengming Liu, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dynamic sparse r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4723–4732, 2022.

Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. *arXiv preprint arXiv:2207.13080*, 2022.

KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5830–5840, 2021.

Prannay Kaul, Weidi Xie, and Andrew Zisserman. Label, verify, correct: A simple few shot object detection method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14237–14247, 2022.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters*, 7(2): 5453–5460, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Sachin Konan, Kevin J Liang, and Li Yin. Extending one-stage detection with open-world proposals. *arXiv preprint arXiv:2201.02302*, 2022.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13619–13627, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022.

Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *arXiv preprint arXiv:2102.09480*, 2021a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021b.

Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3651–3660, 2021.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad 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.

Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 164–173, 2021.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Zhe Lin, Philip Torr, and Jiaya Jia. Open-world entity segmentation. *arXiv preprint arXiv:2107.14228*, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 658–666, 2019.

Kuniaki Saito, Ping Hu, Trevor Darrell, and Kate Saenko. Learning to detect every thing in an open world. *arXiv preprint arXiv:2112.01698*, 2021.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8430–8439, 2019.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020.

Peize Sun, Yi Jiang, Enze Xie, Wenqi Shao, Zehuan Yuan, Changhu Wang, and Ping Luo. What makes for end-to-end object detection? In *International Conference on Machine Learning*, pp. 9934–9944. PMLR, 2021.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9627–9636, 2019.

Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *European conference on computer vision*, pp. 282–298. Springer, 2020.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021.

Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Thang Vu, Hyunjun Jang, Trung X Pham, and Chang Yoo. Cascade rpn: Delving into high-quality region proposal network with adaptive convolution. *Advances in neural information processing systems*, 32, 2019.

Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2965–2974, 2019.

Weiyao Wang, Matt Feiszli, Heng Wang, and Du 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, 2021a.

Weiyao Wang, Matt Feiszli, Heng Wang, Jitendra Malik, and Du 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*, pp. 4422–4432, 2022a.

Wen Wang, Jing Zhang, Yang Cao, Yongliang Shen, and Dacheng Tao. Towards data-efficient detection transformers. *arXiv preprint arXiv:2203.09507*, 2022b.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 568–578, 2021b.

Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33:17721–17732, 2020.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3024–3033, 2021c.

Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. *arXiv preprint arXiv:2207.10661*, 2022.

Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8392–8401, 2021.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10687–10698, 2020.

Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3060–3069, 2021.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14393–14402, 2021.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pp. 391–405. Springer, 2014.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in neural information processing systems*, 33: 3833–3845, 2020.

In the appendix, we first give a comprehensive review of the related works. Then, we provide the detailed description of implementation details and teacher-student learning in Sec. B and Sec. C, respectively. More experimental results are presented in Sec. D, including the VOC to UVO generalization and two ablation studies. Finally, we visualize the score distributions of different methods and show some visualization examples in Sec. E. We promise that *codes and models will be released*.

A More Related Works

Class-agnostic Proposals. The goal of generating object proposals is to locate all the instances in the image regardless of their categories. Before the era of deep learning, the early works (Uijlings et al., 2013; Arbeláez et al., 2014; Zitnick & Dollár, 2014) mainly rely on the hand-crafted clues extracted from images, *e.g.*, edges, texture and colors. Subsequently, the learning-based methods have greatly boosted the performance than the classical algorithm with the localization supervision. The representative work Region Proposal Network (RPN) (Ren et al., 2015) and its variants (Wang et al., 2019; Vu et al., 2019) have been widely used as the prerequisite component to provide the high-recall proposal candidates for the downstream modules (Zareian et al., 2021; Gu et al., 2021). Although these methods prove to be superior in the close-world setup, they tend to overfit to the in-distribution (ID) objects and fail to locate the out-of-distribution (OOD) objects. Joseph et al. (2021) proposes the unknown-aware RPN which automatically labels the potential objects as ground-truth during training and improves the generalization of RPN.

Open-world Instance Recognition. The main obstacle for applying current close-world models is that they treat the unannotated objects as *background* during training and fails to distinguish the *novel* objects from backdrop during inference. To mitigate the issue, Kim et al. (2022) propose the classification-free Object Localization Network (OLN) that replaces the classification head with localization quality head. Although the *novel* objects will not be suppressed, the network merely produces the high-scoring proposals. LDET (Saito et al., 2021) attempts to solve the problem from the perspective of synthesizing images as training source. Specifically, they propose the BackErase data augmentation, which pastes the annotated objects on a background image sampled from a small region so that the unlabeled regions do not contain any hidden objects. LDET further presents a hybrid training strategy to reduce the domain gap between real and synthesized images. The recent work GGN (Wang et al., 2022a) proposes to solve the challenge by adopting a two-stage framework. It first trains a pairwise affinity predictor to extract the semantic object boundaries and generate pseudo labels using the classical grouping algorithms (Arbelaez, 2006; Arbelaez et al., 2010; Shi & Malik, 2000). Finally, Mask-RCNN (He et al., 2017) is trained with the augmented annotations. Although it achieves the promising performance, it suffers from the time-consuming grouping process and thus can not be flexibly used.

B More Implementation Details

We implement our method using the `detectron2` codebase. The optimizer is Adam (Kingma & Ba, 2014) with the base learning of 2×10^{-4} and weight decay of 1×10^{-4}. All the models are trained on 8 NVIDIA A100 GPUs, with 2 samples per GPU. SWORD is trained for 80,000 iterations with the learning rate decaying at the 60,000-th iteration. The backbone is initialized with ImageNet (Deng et al., 2009) pretrained except for the VOC to non-VOC setting. And to ensure a high recall, the number of object queries is set to 2000 for VOC to non-VOC setting and 1000 for the rest settings. In the teacher-student learning process, all the models are trained with the standard `1x` schedule and the number of object queries is 300. In all our settings, we set the value of NMS as 0.7 to remove the redundant predictions. Note that we do not apply the test-time augmentation (TTA) during inference, which proves to be highly effective in the previous works (Wang et al., 2022a).

C More Description about Teacher-student Learning

Teacher-student Learning Details. In the teacher-student learning process, we expect the open-world knowledge of SWORD could be transferred to the close-world student model. In our work, we adopt the standard Deformable-DETR (Zhu et al., 2020) as the student model. As illustrated

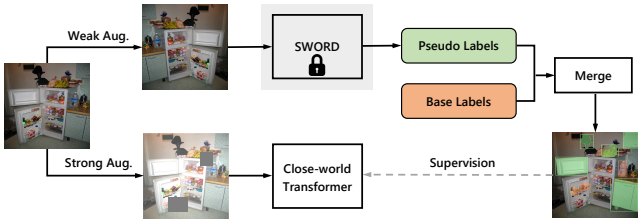


Figure 5: The pipeline of teacher-student learning process. The pre-trained SWORD is first adopted to generate the pseudo labels. Then the original annotations (orange) and generated pseudo ground-truths (green) are merged as supervision to train the close-world Transformer.

Table 9: Comparison of state-of-the-art performance on VOC to UVO setting.

Method	Novel						All					
	AP^b	AR^b_{10}	AR^b_{100}	AP^m	AR^m_{10}	AR^m_{100}	AP^b	AR^b_{10}	AR^b_{100}	AP^m	AR^m_{10}	AR^m_{100}
LDET	9.3	16.0	31.9	4.9	12.3	25.2	22.7	28.1	43.3	18.7	23.9	36.0
Deformable-DETR	7.2	13.5	33.5	3.4	9.5	25.3	23.4	29.4	49.8	19.1	24.0	39.4
SWORD	11.2	16.8	43.1	6.1	13.3	34.9	**24.9**	30.6	55.3	19.6	25.3	45.2
SWORD*	**11.8**	**18.4**	**45.6**	**8.4**	**16.8**	**38.1**	23.4	**31.1**	**59.2**	**21.0**	**28.4**	**49.5**

in Figure 5, the teacher model and student model are given the weakly and strongly augmented images, respectively. The new annotations are merged from the original annotations and pseudo ground-truths generated by the teacher model. Lastly, the student model is trained under the supervision of augmented annotations. Notably, we empirically find that using the IoU scores as proposal confidence for the teacher model leads to better learning results. And to increase the reliability of pseudo labels, the merge step should also be carefully-designed. Specifically, we first use an aggressive NMS value (*e.g.*, 0.3) for the teacher model to remove most predictions. Considering that the pseudo labels should focus on covering the *novel* objects, we discard those proposals having the box IoU with *base* objects higher than 0.5. Finally, the top-k predictions are kept as pseudo ground-truths.

Data Augmentation. Data augmentation has been demonstrated to play an important role in the self-training (Xie et al., 2020; Kaul et al., 2022; Zoph et al., 2020) and semi-supervised methods (Sohn et al., 2020; Liu et al., 2021a; Xu et al., 2021). Following Liu et al. (2021a), we use the random horizontal flip for weak augmentation. And the strong augmentation includes random color jittering, grayscale, Gaussian blur and random cutout operations (DeVries & Taylor, 2017).

D More Experimental Results

Generalization from VOC to UVO. We consider the more challenging cross-dataset generalization where the training source is COCO dataset (Lin et al., 2014) with the partial annotations for VOC categories. Considering that COCO (Lin et al., 2014) is almost exhaustively annotated, the COCO to UVO evaluation actually does not need the model to consider the critical problem of finding *unknown* objects in the unlabeled regions. While the VOC to UVO setup examines the model's ability to discover *novel* objects and generalize to new domains to the greatest extent. Encouragingly, we can see that SWORD and the extended SWORD* both advance the previous best by a large margin in Table 9. For example, SWORD* significantly surpasses Deformable-DETR by 12.1% on AP^b_{100} for *novel* objects and 9.4% on AP^b_{100} for *all* objects. This phenomenon firmly demonstrated that our method has strong capacity for recalling objects in the open-world.

The Effect of EMA Rate. The momentum update of the contrastive head can improve the consistency of the universal object queue. And a larger EMA rate allows the slower feature change. In Table 10, we present the experimental results with various EMA rate α from 0.5 to 0.9999. As illustrated in the first row, with the EMA rate of 0.5, the model gets relatively low results in both AP and AR metrics. This indicates that the model suffers from the detrimental effect of quick transformation of the object center. And the performance is greatly boosted with the EMA rate increases, *e.g.*, the AP^b on *all* objects achieves 6.9% gain by increasing α from 0.5 to 0.9. We observe that the performance becomes stabled when a larger EMA rate (*e.g.*, $\alpha = 0.999$) is applied.

The Effect of Strong Augmentation in Teacher-student Learning. To demonstrate the effectiveness of strong augmentation in teacher-student learning, we ablate the experiments on COCO to

Table 10: **Ablation on the EMA rate**. The results are based on the COCO to UVO setting.

EMA	Novel						All					
	AP^b	AR^b_{10}	AR^b_{100}	AP^m	AR^m_{10}	AR^m_{100}	AP^b	AR^b_{10}	AR^b_{100}	AP^m	AR^m_{10}	AR^m_{100}
0.5	12.7	18.9	39.6	8.9	16.3	27.8	21.1	29.6	51.7	16.9	24.4	35.8
0.9	15.5	21.7	43.5	11.3	19.2	37.4	28.0	34.1	56.2	24.3	30.4	47.8
0.99	15.4	21.0	45.0	11.2	19.0	38.5	28.8	33.8	57.6	25.3	30.6	48.9
0.999	17.5	22.2	48.1	12.8	19.4	40.6	32.0	36.5	61.2	28.0	32.4	51.5
0.9999	17.5	21.8	48.4	11.9	18.6	40.7	32.9	37.0	61.6	28.4	32.7	52.0

Table 11: **Ablation on strong augmentation in teacher-studnet learning**. We evaluate the models on COCO to UVO and VOC to non-VOC settings. And the results are reported on the *novel* objects.

Strong Aug.	**COCO to UVO**						**VOC to non-VOC**					
	AP^b	AR^b_{10}	AR^b_{100}	AP^m	AR^m_{10}	AR^m_{100}	AP^b	AR^b_{10}	AR^b_{100}	AP^m	AR^m_{10}	AR^m_{100}
✗	16.8	22.8	49.0	12.1	20.4	40.8	5.6	21.1	37.9	5.1	19.5	33.1
✓	16.7	22.6	49.8	12.8	20.8	42.4	6.3	21.7	39.6	5.6	20.0	34.5

UVO and VOC to non-VOC settings, respectively. By comparing the two rows in Table 11, it is observed that the model could obtain better performance with the help of strong augmentation. Besides, we observe that the benefit of strong augmentation is more clear on the VOC to non-VOC than the COCO to UVO setup. The reason may attribute to that the annotation density and class diversity of VOC are more limited, which highlights the importance of augmentation.

E Visualization Results

Score Distributions. We visualize the score distributions of different methods on VOC to non-VOC setting in Figure 6. Deformable-DETR (Zhu et al., 2020) can only find out the in-taxonomy objects and thus its socre distribution is mainly located on the low-scoring areas. OLN (Kim et al., 2022) is trained with the positive samples, making it merely produce the high-scoring proposals. Despite it reveals certain open-world ability to locate the *novel* objects, the network can not effectively discriminate the objects and background. As shown in the figure, the scores of OLN outputs are concentrated around 0.6. As contrast to theirs, the proposals of SWORD are able to locate all the objects with more reasonable scores. SWORD not only displays the favorable open-world generalization but also provides distinct confidences for objects and background. Moreover, we could see that the score distribution of SWORD* is further to the right than Deformable-DETR. This indicates that SWORD* is able to detect *novel* objects by learning the knowledge from SWORD, even though it shares exactly the same architecture with Deformable-DETR.

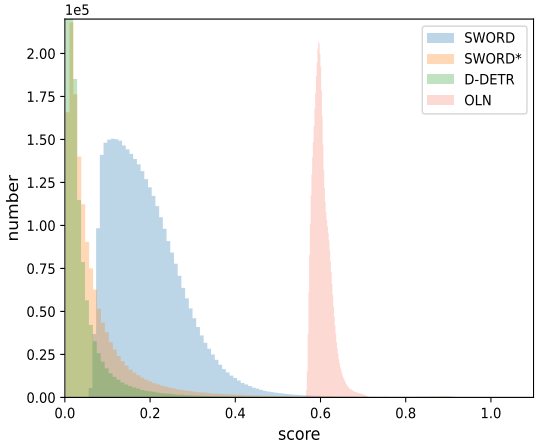


Figure 6: Score distributions of different methods on COCO (Lin et al., 2014) dataset. 'D-DETR' represents Deformable-DETR (Zhu et al., 2020).

Visualization Examples. To showcase the superiority of the proposed SWORD, we further compare the visualization results of different methods in Figure 7. Deformable-DETR (Zhu et al., 2020) is designed for the close-world instance recognition, so it is unable to discover the out-of-taxonomy objects. While OLN (Kim et al., 2022) has demonstrated the open-world ability to locate the *novel* objects, it will also produces numerous false positive predictions, *e.g.*, part of the elephants in the third row and background areas in the fifth row. It could easily see that the proposed SWORD* predicts more accurate and exhaustive segmentation masks. Moreover, it can even find out the unannotated or missing-annotation objects in the ground-truths, *e.g.*, the lamps in the second last row and the kite held by the man in the last example.

(a) Deformable-DETR (b) OLN (c) SWORD* (ours) (d) Ground-truth

Figure 7: Visualization results on VOC to non-VOC setting. The score thresholds for visualization are set as 0.45, 0.65 and 0.45 for Deformable-DETR (Zhu et al., 2020), OLN (Kim et al., 2022) and SWORD*, respectively. It is observed that Deformable-DETR is unable to segment the *novel* objects and OLN produces many false positive predictions. Our model obviously provides the accurate the exhaustive segmentation masks.