Robot Instance Segmentation with Few Annotations for Grasping

Anonymous ECCV 2024 Submission

Paper ID #3

Abstract. The ability of robots to manipulate objects relies heavily on their aptitude for visual perception. In domains characterized by cluttered scenes and high object variability, most methods call for vast labeled datasets, laboriously hand-annotated, with the aim of training capable models. Once deployed, the challenge of generalizing to unfamiliar objects implies that the model must evolve alongside its domain. To address this, we propose a novel framework that combines Semi-Supervised Learning (SSL) with Learning Through Interaction (LTI), allowing a model to learn by observing scene alterations and leverage visual consistency despite temporal gaps without requiring curated data of interaction sequences. As a result, our approach exploits partially annotated data through self-supervision and incorporates temporal context using pseudo-sequences generated from unlabeled still images. We validate our method on two common benchmarks, ARMBench [34] mix-object-tote and OCID [37], where it achieves state-of-the-art performance. Notably, on ARMBench, we attain an AP_{50} of 86.37, almost a 20% improvement over existing work, and obtain remarkable results in scenarios with extremely low annotation, achieving an AP_{50} score of 84.89 with just 1% of annotated data compared to 72 presented in [34] on the fully annotated counterpart [1] .

1 Introduction

Acquiring accurate instance segmentation masks requires training a model on vast amounts of data with high-quality pixel-level annotations. While collecting raw sensory data (images) is relatively easy, annotating object instance masks down to individual pixels becomes prohibitively expensive when scaling up perception tasks. As a result, models trained on limited annotated data inevitably face challenges when deployed in the real world due to domain variation and evolving environments. This problem is central in robotics, where robots rely on spatial perception extracted from sensory inputs.

To use large amounts of unlabeled data, **Semi-Supervised Learning (SSL)** assumes that only a portion of the data is labeled: either a subset of observed scenes or some objects within each scene. The model then uses its own predictions as pseudo-labels to extract learning signals from the remaining unlabeled data [36,39,48,50,52]. Therefore, a model attempting to learn from its own noisy labels early in training may stagnate rather than generalize.

[1] The code will be made available upon publication.

Looking beyond spatial cues of still images, video sequences contain temporal information that a model can exploit to enforce consistency across frames and improve generalization. Recent advancements focusing on **Learning Through Interaction (LTI)** highlight the significance of providing the model with temporal perception. LTI enables the model to peer into the underlying dynamics of its domain by observing actions and their consequences [8, 21, 38]. The leading approaches entail observing a scene that undergoes various changes, such as objects being placed or extracted. By constructing the data in the form of "before" and "after" sequences [24, 27, 43, 44, 51], localized changes in illumination, deformation, and articulation of objects allow the model to refine its interpretation of the environment. The leading LTI techniques either prescribe multi-stage training that pre-trains on specialized datasets [27] or restrict input observations to strictly gradual changes at small time intervals [2]. Interestingly, leading methods for Video Image Segmentation regularly resolve long image sequences depicting instances popping in and out of view [16, 43, 44, 51]. These incorporate a reassociation loss to overcome changes of occlusion, instance pose and appearance. However, learning from videos relies either on significant investment in manual annotation of every object in every video frame or on video frames occurring at sufficiently small time intervals. Our main insight is that although each paradigm compensates for the weakness of the other (SSL lessens the annotation effort while LTI leverages temporal information), naively combining the two amplifies their drawbacks—reinforcing noisy labels across entire sequences. In this work, we propose a solution in the form of a novel framework that incorporates the learning paradigms of SSL and LTI to enhance performance in the few-annotations scenario, in which only a tiny fraction of the dataset is annotated (and the rest is unlabeled). Our method simultaneously addresses the challenges of LTI and SSL. We eliminate the need for specialized datasets required for LTI by using pseudo-sequences generated from still images to mimic scene interaction. We also overcome the main obstacles to SSL by preventing noisy self-predictions from obscuring the learning signal through coupling prediction heads, thus stabilizing predictions early in training. Our framework is model-agnostic, complementing

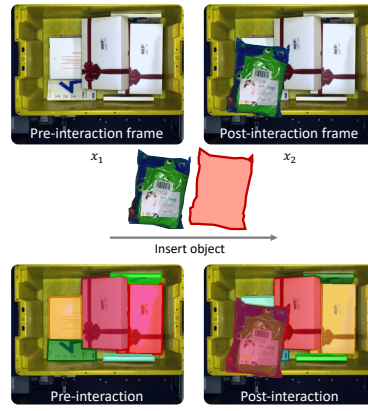


Fig. 1: Method: Pseudo-sequence generation from a single unlabeled image. The input is weakly augmented to produce the "before" image x_1 and augmented again to yield the "after" image x_2. To emulate scene interaction, objects are drawn from the object memory bank, transformed, and inserted into the "after" frame. The segmentation model's task is to simultaneously associate objects that persist between frames (subject to occlusion), maintain consistency of object instance embedding, and correctly predict the ground-truth mask of the added objects.

existing (and future) segmentation models with temporal perception through end-to-end training. Additionally, we propose an automated pseudo-label criteria that discards low-quality predictions.

The resulting framework can be considered the first to employ self-supervised learning through interaction, achieving better performance than each paradigm individually. We set a new state-of-the-art on the ARMBench [34] benchmark and OCID [37] (RGB only). Notably, our method trained on 1% of annotated data surpasses the performance of the well-established Deformable DETR [53] architecture, even when trained on 10× additional annotated data (improving +16.86 AP using Swin-L Transformer as feature extractor).

2 Background and Related Work

Of the various approaches to instance segmentation, we are interested in those that excel without full supervision [8, 21]. This section provides an overview of relevant works on partial supervision and learning from sequences.

Partial supervision methods use the few annotated examples available (if any) and maintain consistent predictions for similar objects in the scene [57]. In recent years, most efforts focused on contrastive learning that extracts embedding from object instances and aims to bring same-class embedding closer while pushing other classes further apart [4, 36]. That said, progress in object classification and detection does not readily carry over to image segmentation, where the effectiveness of self-supervision lags behind full-supervision in challenging domains of cluttered objects with many occlusions [54]. Unsurprisingly, these domains are also more complicated for humans to annotate.

Scene modulation is a concept that aims to extract additional learning signal by familiarizing the model with objects that undergo gradual alterations within a scene [7, 24, 41, 42, 47] where objects are viewed in many configurations, as well as different clutter and lighting conditions. This offers a substantial advantage in detecting and identifying objects that may deform or exhibit variations, thereby enhancing the robustness of the segmentation. Note, however, that assembling large dedicated datasets of objects is resource-intensive and challenging to apply effectively to new scenes featuring previously unseen objects. A recent work [47] achieved significant improvement by incorporating simulated data before transferring to real world scenes [15, 24]. The main drawback of using synthetic data is the high cost of creating photorealistic rendering that accurately captures the physical properties of every object in the scene. Often times this results in idiosyncrasies that are picked up by the model and become a source of error when encountering real world data.

Frame sequences offer additional information along the time dimension. As with scene modulation, the model learns to recognize and identify related instances throughout a series of images [2]. Recent advancements in video instance segmentation (VIS) methods, exemplified by SeqFormer [43] and IDOL [44], leverage sequential consistency of instances for online object segmentation and tracking. They employ contrastive loss to ensure that instance representations

are distinguishable from other instances in the same frame and over previous frames. In CTVIS [51] the model also taps into future frames.

Learning through interaction pushes the notion of sequences even further by specializing in image sequences that depict predefined and controlled scene manipulation. Consecutive frames in these meticulously assembled datasets exhibit large temporal gaps, unlike video data, and changes are usually confined to localized actions on few object instances [27]. This locality constraint persists through frame sequences, allowing LTI approaches to infer which instances have actually changed and which are merely affected by variations in lighting, occlusion and deformation, as a result of the action performed. The model quickly learns to segment an object that is added or removed, using a few hundred labeled image pairs. Since assembling such specialized datasets requires significant effort, the next stage in training artificially inserts cropped instances from high-confidence mask predictions into unlabeled still-images to emulate interactions. In STOW [24], the model is additionally trained on synthesized virtual scenes and then evaluated on real-world data.

The above advancements present an interesting question: In real-world applications where the model inevitably encounters a changing domain, is it possible to continuously learn (post deployment) without supervision by leveraging the temporal information of video sequences using the causal awareness of LTI? Importantly, can this be achieved without investing in a proprietary dataset or reliance on a specific instance segmentation model?

3 RISE

We introduce a novel framework called **R**obot **I**nstance **Se**gmentation for Few-Annotation Grasping (**RISE**) that unifies learning from temporal signals (through interactions) and spatial signals (through self-supervision). RISE is trained end-to-end on still images, and enables self-supervision to learn from temporal consistency when scene objects are moved, added or removed. Because of this, RISE does not require a meticulously compiled dataset of before and after image pairs of scene interaction, nor does it require a large dataset of labeled instances — thus it is more readily capable of handling domain variations that commonly occur in the real world.

The framework accommodates both supervised and semi-supervised data, comprising an instance segmentation model (Sec. 3.1) that is enhanced to extract learning signals from both instance-association and consistency losses (Sec. 3.2).

3.1 Instance Segmentation

Object instance segmentation begins with the input image x which first undergoes feature extraction by an extractor backbone. The features are then fed into an instance level embedding encoder that outputs 300 tokens that serve the decoder, which emits instance embedding z_i into the predictions heads for class, bounding box and mask of object instance i. In this work we evaluate various

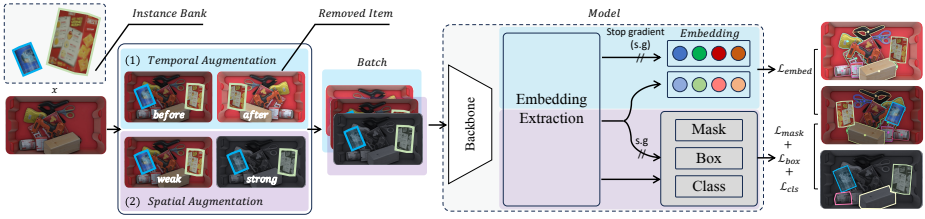


Fig. 2: RISE framework (from left to right). Given an unlabeled image x and a bank of known instances, we perform (1) temporal (blue) and (2) spatial (purple) augmentations. Temporal augmentation adds weak augmentation and inserts K instances from the bank to create x_1 (the "before" frame). Another round of weak augmentations, combined with adding/moving/removing a subset of the K added instances, produces x_2 (the "after" frame). Spatial augmentations adds strong augmentations to create x_3. The three images are batched and fed into the model, where the backbone extracts features that are then encoded into instance embedding. The instance embedding from x_1 and x_2 are used to compute $\mathcal{L}_{embed}$ Eq. (4). The embedding from x_1 serve as pseudo-labels against the embedding from x_3 in the self-supervised loss $\mathcal{L}_{\text{u}}$ Eq. (5).

established backbones: Resnet50, Resnet101 [14], and Swin-L [28] as options for the feature extractor. As embedding decoder we chose Deformable DETR [53] as a strong spatial decoder for its ability to learn object queries as features. The prediction heads for class labels and box coordinates are feed-forward networks (FFNs), whereas the mask prediction head is a Feature-Pyramid network (FPN) [25] that uses multi-scale features from the decoder's last layers, followed by an FFN whose output mask is scaled up to match the original image size.

When the input is also accompanied by labels $\mathbf{y}$, the supervised component of the loss constitutes a class label loss $\mathcal{L}_{cls}$; $\mathcal{L}_{box}$ that combines L_1 loss and generalized Intersection over Union (gIoU) loss [35]; $\mathcal{L}_{mask}$ as the sum of the Dice loss [33] and Focal loss [26]:

$$\mathcal{L}_{\text{s}} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{box} + \lambda_2 \mathcal{L}_{mask}, \tag{1}$$

where λ_1 and λ_2 are the loss coefficients.

Recent advancements in object detection incorporated optimal transport (OT) to address the optimal assignment between predictions and ground truth, as proposed by Ge et al. [9] and YOLOX [10]. Therefore, we compute the pairwise cost between predictions and ground truth instances, determining the optimal assignment for the top k predictions associated with each instance.

While operating on labeled data, object instances that are successfully segmented by the model are stored in a memory bank [11]. This object bank will be used in the self-supervised phase, during which labeled instances are randomly selected and augmented to emulate scene interaction.

3.2 Learning Through Interaction

Observing interactions shares commonality with Object Tracking, which goes beyond traditional object detection. It leverages discriminative representation

of instances across frames and of different instances belonging to the same class. The resulting representation is more robust to occlusion and identity switches, as demonstrated in [22, 45].

Given an unlabeled input frame x containing an unknown number of object instances, we introduce a new augmentation strategy to create a pair of pre- and post-interaction frames. The first, pre-interaction frame $x_1 = \psi_1(x)$ is an augmentation ψ_1 of x where we stochastically insert K objects from a bank of known instances. Each of the K objects is also individually augmented (e.g., scale, position, rotation, flip, color). The second, post-interaction frame $x_2 = \psi_2(x_1)$ is an augmentation ψ_2 of x_1 where we also remove several of the objects added to x_1 or insert a few more objects from the instance bank (with augmentations). Note that both x_1 and x_2 are spatially augmented with rotation, crop, and scale to convey a sense of motion to the observer (inspired by SeqFormer [43]).

Augmentation Strategy Inserting new objects into a dense scene may lead to significant occlusions and even conceal the objects we intend to learn. Therefore, we devise a strategy that randomizes labeled objects from the instance bank and distributes them preferentially around the periphery of the frame:

$$(u, v) = Beta(\alpha, \beta) \cdot [w, h] \tag{2}$$

where (u, v) is the top-left corner where the object is inserted, drawn from distribution $Beta(\alpha, \beta) \in \mathbb{R}^2$, and w, h are the feasible horizontal and vertical regions that ensure that the object is contained within the frame (see Appendix A). The choice of $Beta$ and its parameters reduces the likelihood of objects inserted near the center, where they might obstruct unlabeled objects. Another safeguard prevents inserting an object if it would overlap with any of the previously inserted objects by more than 85%.

Association Loss The resulting frames x_1, x_2 contain a total of N and M object instances, respectively. Importantly, the objects' small projective and illumination transformations compel the model to learn robust representations that maintain consistency for occurrences of the same instance in a changing scene (illustrated in Fig. 2). Each embedding $i \in N$ extracted from the first frame x_1 is matched against every embedding $j \in M$ in the second frame x_2, forming the association score $f(i, j)$ between instance i and instance j:

$$f(i,j) = \frac{1}{2}\left[\frac{\exp(z_j^T \cdot z_i)}{\sum_{k=1}^{M}\exp(z_k^T \cdot z_i)} + \frac{\exp(z_j^T \cdot z_i)}{\sum_{k=1}^{N}\exp(z_j^T \cdot z_k)}\right] \tag{3}$$

We consider the embedding of $\hat{j} = \arg\max f(i, j)$ as a positive example for the given instance i if $f(i, \hat{j}) > 0.5$, otherwise it is considered a negative example. We employ an embedding contrastive loss [3] to learn object representation from the observed interaction frames:

$$\mathcal{L}_{embed} = -\log \frac{\exp(z_i \cdot z_j^+)}{\exp(z_i \cdot z_j^+) + \sum_{z_i^-}\exp(z_i \cdot z_j^-)} \tag{4}$$

where z_i is the embedding of instance i in the first frame, z_j^+ is the embedding of the instance j in the second frame that ideally represents the same instance, and z_j^- are the embedding of the remaining instances (negative views). The loss $\mathcal{L}_{embed}$ pulls same-instance embedding closer together while pushing apart representations of different instances.

3.3 Self-Supervision

To better leverage spatial information in unlabeled data, we employ the segmentation model (Sec. 3.1) toward Semi-Supervised Learning (SSL). Inspired by [36], we include a consistency regularization loss and extend it to accept unlabeled images alongside labeled objects inserted from the instance bank.

Recall that $x_1 = \psi_1(x)$ is a weak augmentation of the unlabeled input image x. In this context, we'll denote $x_w = x_1$. We apply another round of weak augmentations to x_w, followed by a strong augmentation ϕ to produce $x_s = x_3 = \phi(x_1)$. The strong augmentations comprise Color jitter, Planckian jitter [55], Gaussian blur, and gray-scale that are applied via RandAugment [6]. We feed both x_w and x_s into the model. Class labels, bounding boxes, and segmentation masks for weakly augmented inputs x_w are treated as pseudo-label targets (in the absence of ground truth) that are compared against the model's prediction on x_s. The unsupervised consistency regularization loss:

$$\mathcal{L}_{\text{u}} = \hat{\mathcal{L}}_{cls} + \lambda_1 \hat{\mathcal{L}}_{box} + \lambda_2 \hat{\mathcal{L}}_{mask} \tag{5}$$

It is similar to the supervised loss $\mathcal{L}_{\text{s}}$ (Eq. (1)), with the distinction that pseudo-labels are used in place of ground-truth labels. Gradients are not computed during the forward pass of x_w (as illustrated in Fig. 2) as it constitutes the ground truth. We introduce the following refinements to stabilize the model during self-supervised training.

Refined Consistency Learning It is common practice to filter out pseudo-labels with low prediction scores in order to reduce the model's exposure to errors during self-supervised training. The filters are often thresholds or quantiles that are either fixed, dynamic, or scheduled [18, 40]. Thresholds, by nature, are more restrictive, discarding all predictions below their stated value. However, during the early stages of self-supervision, the model may emit most of its predictions slightly below the threshold, resulting in very few labels contributing towards learning. On the other hand, quantiles ignore the scores entirely and allow any prediction, provided that its score meets the rank requirement of the quantile. Because most models output a fixed number of predictions to accommodate crowded scenes (regularly exceeding 300 predictions), a quantile may become too lenient and include low-score predictions of poor quality, potentially degrading the model's performance as training progresses.

In Appendix D, we demonstrate that early in training, setting the quantile too low lets in more predictions of low-quality signals, interfering with the model. Alternatively, setting the bar (too) high [36] risks missing out on meaningful supervision signals.

To reconcile the limitations of both thresholds and quantiles, we propose a cascade approach. First, a more relaxed class threshold γ_t^{cls} removes instances whose class scores $\hat{c}_i \in \hat{\mathbf{c}}$ are deemed unusable, followed by a quantile selection Q [40] of the leading predictions. The resulting class pseudo-labels $\hat{\mathbf{y}}$ is given by:

$$\hat{\mathbf{y}} = Q(\ \hat{\mathbf{c}} > \gamma_t^{cls}\ ;\ p_t). \tag{6}$$

The threshold γ_t^{cls} discards instances with class scores $\hat{c}_i$ below it and tightens over time (training steps t). Conversely, the quantile $Q(p_t)$ loosens over time with its probability $p_t = 0.995 \cdot (1 - t/T)$ decays over subsequent training steps t, with T denoting the total number of training steps. As a result, the quantile allows more predictions into the model as training progresses. This strategy can mitigate incorrect model beliefs and reduce confirmation biases. We evaluate this strategy quantitatively and demonstrate its advantage over thresholds and quantiles in Tab. 4, with additional details in Appendix C. We recognize that exploring different quantile strategies may further improve self-supervision and set it aside for future work.

Coupled Prediction Heads The standard approach to filtering pseudo-mask predictions employs a pixel-wise confidence threshold γ_t^{mask} that is applied to each pixel (u, v) of instance mask $\hat{m}_i$:

$$\hat{m}_i^{u,v} = \begin{cases} 1 & \text{if } h^{\text{mask}}(z_i^w) > \gamma_t^{\text{mask}}, \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where h^{mask} is the mask head output for instance embedding z_i^w obtained from the weakly augmented frame x^w.

Unlike masks, the prediction quality of bounding boxes is less correlated with high label scores. As such, recent SSL methods for object detection employ multiple passes to refine box predictions [1, 49]. Interestingly, we observe that the model learns to predict high quality masks well before it effectively predicts bounding boxes. Thus, we propose a coupling of the mask and box prediction heads so that during training, pseudo-boxes $\hat{b}_i$ are obtained by bounding their corresponding instance segmentation masks $\hat{m}_i$:

$$\hat{b}_i = \left[\min_u \hat{m}_i\ \min_v \hat{m}_i\ \max_u \hat{m}_i\ \max_v \hat{m}_i\right]. \tag{8}$$

We refer to this simple yet effective technique for pseudo-box assignment as Mask-to-Box (M2B) and demonstrate its advantage over the standard approach to predicting pseudo-boxes in Tabs. 3 and 5.

Multi-Label Matching A common practice in object segmentation is to apply non-maximum suppression (NMS) to eliminate redundant predictions. In our case, instance overlaps are common since objects are inserted at random as part of pseudo-sequence generation. Therefore, to make better use of the model's predictions during training, we introduce a new adaptation of Label-Matching (LM) [1], whereby we retain several overlapping predictions that coincide with the dominant class label (instead of discarding all but one). We call this method Multi-Label Matching (MLM) and conduct an ablation study to assess its contribution to self-supervision (Tabs. 3 and 5), demonstrating its advantage.

3.4 Unified Framework

The complete architecture of RISE is presented in Fig. 2, comprising an LTI branch and an SSL branch that converge into a unified loss:

$$\mathcal{L}_{\text{total}} = \mathbb{1}[\mathbf{y} \neq \varnothing]\mathcal{L}_{\text{s}} + \lambda_3\mathcal{L}_{\text{embed}} + \mathbb{1}[\mathbf{y} = \varnothing]\lambda_4\mathcal{L}_{\text{u}}, \tag{9}$$

where $\mathbb{1}$ indicates that the supervised loss $\mathcal{L}_{\text{s}}$ and unsupervised loss $\mathcal{L}_{\text{u}}$ are used according to the availability of ground-truth labels $\mathbf{y}$, and $\mathcal{L}_{embed}$ denotes the weighted combinations of the association loss. Hyperparameter search for λ_3 and details on λ_1, λ_2 and λ_4 are provided in Appendix A.

4 Experiments

We conduct a series of experiments to evaluate the performance of the proposed approach in the Robotic Item Grasping domain. This domain is of high relevance to automated distribution warehouses, where robotic arms pick and place items inside totes. The experiments target a range of labeled data ratios, meaning that we intentionally restrict the model's access to only a certain portion (%) of the labeled samples, and treat the remaining samples as unlabeled.

4.1 Setup

Datasets Our main focus is the ARMBench [34] mix-tote benchmark comprising 44,234 images, split into 30,992 training images and 6,637 and 6,605 images for validation and testing, respectively. The images are not organized into sequences nor do they describe a localized action. Every object in the scene belongs to a single "object" category and is associated with a manually annotated instance mask.
The OCID [37] dataset (containing 2,390 images and 31 classes) for various rates of labeled-to-unlabeled data, and compare it to the current state-of-the-art [32]. We use the same RISE configuration (e.g., Beta function, thresholds, etc.) for both datasets. The results in Tab. 2 illustrate that our method is readily applied to new datasets without requiring domain-specific configuration adjustments.

Evaluation We evaluate our method using the standard Average Precision (AP). We measure the overall AP across 10 IoU thresholds $[0.50, \ldots, 0.95]$, as well as the IoU thresholded precision AP_{50} and AP_{75}. For OCID we use only the AP_{50} to be consistent with prior art.
In terms of partitioning, we use 100%, 10%, 2%, 1% and 0.5% of the data as fully annotated, and the remaining as unlabeled for ARMBench, and 100%, 10% and 5% for OCID. We compare RISE with the officially reported performance from the ARMBench [34] and RoboLLM [29], in which a model was trained on the entire training set. This baseline is the existing state-of-the-art on the ARMBench dataset. In addition, we compare RISE with Deformable DETR [56] (denoted DeDETR).

4.2 Results

Tab. 1 shows the results of RISE on various data partitions of labeled/unlabeled ratios of the ARMBench, compared with Deformable DETR and SAM [19] (fine-tuned), as well as the results reported by the authors of ARMBench [34]. Both DeDETR and RISE use Swin-L [28] (197M parameters) as backbone, while SAM uses ViT-H [23] (636M parameters) and RoboLLM [29] uses Beit-3 base (87M parameters). Across all partitions, RISE outperforms the other methods. Fig. 3 illustrates high-quality masks predicted by RISE trained on 1% of the labeled, with 99% of the remaining data treated as unlabeled. Most of the line-of-sight objects are accurately segmented and a few heavily occluded objects are missed. Importantly, RISE trained on 1% annotated data performs better than DeDETR and SAM trained on 10% annotated samples (10× the amount of annotations for training/fine-tuning).

Table 1: Comparison between our method and prior art on ARMBench [34] mix-object-tote instance segmentation challenge with subset of annotations. The first column represents the number of annotated samples used for training, with the rest treated as unlabeled data for self-supervised learning. The second column details the method name and the next columns are AP measures, with best performers marked in **bold**.

%		ResNet-50			ResNet-101			ViT		
Labeled	Method	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
0.5%	DeDETR [53]	27.03	29.32	26.65	28.36	30.69	28.14	36.47	36.46	31.75
(155)	M2F [5]							55.3	59.9	54.7
	SAM [20]							61.38	74.04	63.51
	RISE	**66.15**	**78.80**	**69.67**	**71.40**	**82.10**	**72.30**	**72.14**	**83.25**	**73.73**
1%	DeDETR [53]	27.17	29.64	26.67	30.38	34.52	29.73	39.46	39.44	33.51
(309)	YOLACT				36.1	59.2	44.8			
	M2F [5]							58.6	64.7	58.6
	SAM [20]							67.42	82.26	70.93
	RISE	**69.10**	**82.10**	**73.80**	**73.00**	**83.25**	**73.94**	**73.72**	**84.89**	**74.89**
2%	DeDETR [53]	31.79	36.5	31.81	33.14	39.70	34.19	42.15	66.20	43.39
	M2F [5]							61.4	68.5	61.2
(618)	RISE	**72.80**	**82.90**	**74.40**	**73.66**	**83.44**	**75.89**	**73.92**	**84.00**	**76.25**
10%	DeDETR [53]	48.00	57.44	48.17	52.23	59.98	49.8	59.19	75.5	60.42
(3,099)	YOLACT				47.40	68.20	52.70			
	M2F [5]							68.2	76.5	68.2
	SAM [20]							71.47	82.78	73.96
	RISE	**73.39**	**83.48**	**75.09**	**74.27**	**84.33**	**75.54**	**74.95**	**85.16**	**76.26**
100%	ARMBench [34]	-	72.00	61.00	-	-	-	-	-	-
(30,992)	DeDETR [53]	52.11	60.38	52.52	53.80	62.00	52.80	62.75	77.03	63.40
	M2F [5]							73.00	81.2	74.00
	RoboLLM [29]								82.0	67.00
	RISE	**73.41**	**83.53**	**75.15**	**74.47**	**84.74**	**75.93**	**76.04**	**86.37**	**77.51**

Tab. 2 compares the performance on various data partitions of labeled/unlabeled ratios of the OCID dataset, showing the advantage of RISE.

Table 2: Evaluation on OCID [37] (RGB only). $^+$ denotes Second stage networks. Bottom rows show the performance when only a portion (%) of annotations are used. The proposed approach achieves significantly better results even with few annotations compared to prior art.

Method	AP_{50}
UCN [46]	54.8
UCN$^+$+ [46]	59.1
Mask2Former [5]	67.2
MSMFormer [31]	72.9
MSMFormer$^+$ [31]	73.9
MRCNN [13]	77.6
RISE	**78.2**
RISE (5%)	75.1
RISE (10%)	77.3

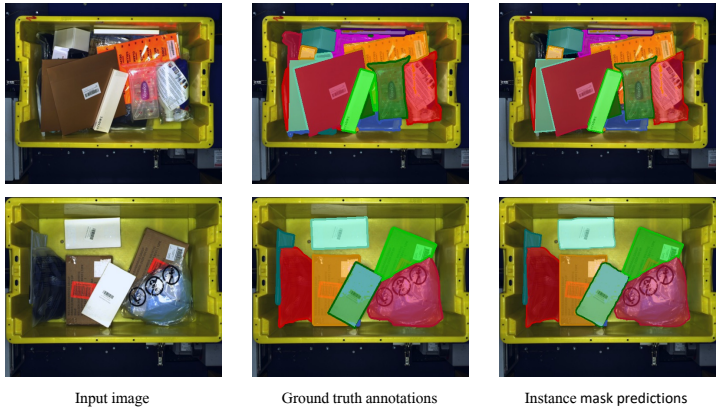


Fig. 3: Qualitative results of RISE using ResNet-101 backbone, trained on 1% of the labeled data (99% treated as unlabeled). Comparing the ground truth (center column) and the predicted masks (right-most column), we see that the majority of large items are accurately segmented, while some of the smaller or heavily occluded objects are occasionally missed. The segmentation masks are continuous, indicating high confidence for every object. Mask boundaries are the only regions with instance–background ambiguity (evident by mild noise at object boundaries).

Foundation Model Comparison

As an additional baseline we compare RISE with the "Segment Anything" (SAM) foundation model [19], fine-tuned on a subset of the ARMBench dataset. In Tab. 1 we demonstrate that despite SAM's unrivalled ability to *segment anything*, it is prone to over-segment and produce mask artifacts, even after fine-tuning on a small portion of domain-specific images. Fig. 4 shows an example where SAM, fine-tuned on 1% of the data still struggles with accurately discerning objects, resulting in fragmented and incomplete object masks and mask

predictions that target less significant elements of the image (such as packaging features, rivets and shadows). The numerous false positive predictions impact the overall performance.

4.3 Ablation Study

We provide an ablation study of the various design choices made in implementing RISE: impact of losses, pseudo-label threshold strategies and parameters. Tab. 3 details the contribution of the different elements within RISE on the fully-supervised training set. The most substantial improvement is attributed to the Pseudo-Sequence (PS) strategy outlined in Sec. 3.2. The coupling of prediction heads in Mask-to-Box (M2B in Eq. (7)) refines the supervision signal for box predictions, further improving the performance. Combining it with Multi-Label Matching (MLM) and Optimal Transport (OT) yields the best performing version of RISE.

Fig. 4: Fine-tuned SAM (1% ARMBench data) showing many fragmented masks and false positive artifacts.

Table 3: Ablation study on different components of RISE. Pseudo-Sequences (PS), Optimal Transport (OT), Mask-to-Box coupling (M2B) and Multi-Label Matching (MLM). We evaluate the contribution of these component using 100% of the annotated data, showing that the combined approach achieves the best results.

PS	OT	M2B	MLM	AP
				53.80
✓				73.85
	✓			62.92
✓	✓			74.16
✓	✓	✓		74.35
✓	✓	✓	✓	**74.47**

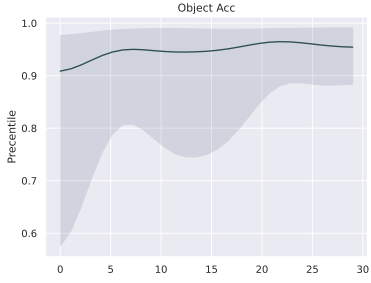
Next we evaluate the pseudo-label elimination strategy of either a standard score threshold or quantile function, compared with the proposed cascade approach (Eq. (6)). Notably, setting the threshold or quantile too low would include more false positive predictions in training. Setting them too high would

Table 4: Ablation study of pseudo-label threshold strategy, comparing the score threshold γ_t^{cls}, the quantile $Q(p_t)$, and the proposed score filtering cascade (Eq. (6)) that applies a threshold followed by a quantile $\gamma_t^{cls} \rightarrow Q(p_t)$ (or reversed $Q(p_t) \rightarrow \gamma_t^{cls}$). The best performance is achieved for the proposed cascade approach.

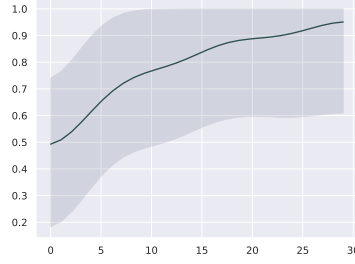
Threshold Strategy	AP	AP_{50}	AP_{75}
Threshold only γ_t^{cls}	72.91	83.39	74.38
Quantile only $Q(p_t)$	72.59	83.2	74.44
Cascade $\gamma_t^{cls} \rightarrow Q(p_t)$	**74.47**	**84.33**	**75.93**
Cascade $Q(p_t) \rightarrow \gamma_t^{cls}$	73.0	82.75	74.55

eliminate correct predictions since very few predictions would meet the required prediction score. This holds even when both threshold and quantile are dynamic (changing via predefined schedule). Tab. 4 shows that the cascade approach which first enforces a lenient threshold and then a quantile yields the best results.

In Fig. 5 we illustrate that setting the threshold too high at any point during training would eliminate many true-positive predictions, solely due to the model predicting a low class label score. This is more prominent early in training, since the model gains confidence in its predictions as training progresses. We visualize the "ideal" threshold that would only retain true-positives and ensure that no false-positive pixels are passed through.



(a) Label accuracy over time The solid-black line represent the "ideal" class threshold γ^{cls} that would eliminate all false-positive predictions and retain only true-positives. The regions highlighted in gray denote the standard-deviation of the ideal threshold.



(b) Mask accuracy over time The solid-black line describes the mask threshold γ^{mask} that for each instance, distinguish between pseudo-masks and background. The gray-highlighted regions denote the standard-deviation .

Fig. 5: Label and Mask Accuracy of pseudo-labels. For both he x-axis measures training steps in multiples of $\times 1000$.

Finally, we measure the impact of the different pseudo-box strategies. Tab. 5 shows the resulting AP for RISE trained on 1% of the labeled data and various values of threshold γ^{mask}. The standard approach filters the boxes by thresh-

Table 5: Ablation study of mask threshold γ^{mask}, class threshold γ^{cls} and pseudo-box strategies, showing AP for **1% annotated data**. The Pseudo-box column corresponds to the standard pseudo-box approach which discards box predictions corresponding to pseudo-labels below the threshold γ^{cls}. The Mask-to-Box (M2B) and the combined Mask-to-Box with Multi-Label-Matching (M2B+MLM), introduced in this work, extract pseudo-boxes from pseudo-masks and instead filter individual pixels whose score fall below γ^{mask}. The table shows that M2B+MLM produced the best results.

γ^{cls} or γ^{mask}	Pseudo-box	M2B	M2B + MLM
0.5	71.98	72.71	**73.00**
0.6	71.70	72.55	**72.87**
0.7	71.57	72.28	**72.86**

olding the class prediction score using γ^{cls}. We denote by M2B the effect of Mask-to-Box (extracting bounding boxes from the predicted instance masks), and denote by MLM the use of multiple boxes towards the self-supervised loss $\mathcal{L}_{\text{u}}$ in Eq. (5). The results demonstrate that employing a mask threshold of $\gamma^{mask} = 0.5$ in conjunction with M2B + MLM achieves the best performance.

5 Conclusion

In this work, we present RISE, a novel framework that incorporates semi-supervised learning with learning through scene interaction in the context of a few-annotation data regime. RISE is modular and can complement other segmentation models that emit intermediate instance embedding. We demonstrate that RISE improves AP_{50} by over +10 compared to previous state-of-the-art, after training end-to-end on just 0.5% of the labeled data (with 99.5% of the data treated as unlabeled). With just 1% of the labeled data, RISE achieves better performance than the baselines (DeDETR, SAM, RoboLLM) trained on 10× the amount of labeled data. On OCID (RGB), RISE sets a new state-of-the-art, and is near state-of-the-art when restricted to a fraction of the annotations. For future work, we intend on leveraging "before" and "after" observations directly using robotic item grasping in real-world environments (rather than synthetically inserting instances into images), with the overarching goal of lifelong learning for robot perception.

A limitation of the proposed approach is that it underperforms when presented with objects that make few or no appearances in the truncated (labeled) training data. Access to a handful of annotated examples means that not all objects are encountered during training, resulting in some cases where two objects are segmented as one. In the context of robotic grasping, this may lead to a failed object-grasping attempt. However, since grasp failures also alter the scene, we believe that capturing snapshots of the scene before and after the attempted interaction would help improve the grasping precision in the long term.

References

1. Chen, B., Chen, W., Yang, S., Xuan, Y., JieSong, Xie, D., Pu, S., Song, M., Zhuang., Y.: Label matching semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 8, 3
2. Chen, H., Venkatesh, R., Friedman, Y., Wu, J., Tenenbaum, J.B., Yamins, D.L., Bear, D.M.: Unsupervised segmentation in real-world images via spelke object inference. In: European Conference on Computer Vision. pp. 719–735. Springer (2022) 2, 3
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1597–1607. PMLR (Jul 2020), https://proceedings.mlr.press/v119/chen20j.html 6
4. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 22243–22255. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/hash/fcbc95ccdd551da181207c0c1400c655-Abstract.html 3
5. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation (2022) 10, 11
6. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.: RandAugment: practical automated data augmentation with a reduced search space. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 18613–18624. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html 7
7. Dwibedi, D., Misra, I., Hebert, M.: Cut, paste and learn: Surprisingly easy synthesis for instance detection. In: Proceedings of the IEEE international conference on computer vision. pp. 1301–1310 (2017) 3, 2
8. Garg, S., Sunderhauf, N., Dayoub, F., Morrison, D., Cosgun, A., Carneiro, G., Wu, Q., Chin, T.J., Reid, I.D., Gould, S., Corke, P., Milford, M.: Semantics for robotic mapping, perception and interaction: A survey. ArXiv **abs/2101.00443** (2021) 2, 3
9. Ge, Z., Liu, S., Li, Z., Yoshie, O., Sun, J.: Ota: Optimal transport assignment for object detection. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 303–312 (2021) 5, 3
10. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021) 5
11. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. arXiv preprint arXiv:2012.07177 (2020) 5
12. Grad, E., Kimhi, M., Halika, L., Baskin, C.: Benchmarking label noise in instance segmentation: Spatial noise matters (2024) 5
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask r-cnn. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 2980–2988 (2017) 11
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jun

2016), https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html 5, 1
15. Horváth, D., Bocsi, K., Erdös, G., Istenes, Z.: Sim2real grasp pose estimation for adaptive robotic applications. ArXiv **abs/2211.01048** (2022) 3
16. Ke, L., Danelljan, M., Ding, H., Tai, Y.W., Tang, C.K., Yu, F.: Mask-free video instance segmentation. In: CVPR (2023) 2
17. Kim, B., Choo, J., Kwon, Y.D., Joe, S., Min, S., Gwon, Y.: SelfMatch: combining contrastive self-supervision and consistency for semi-supervised learning. arXiv preprint (Jan 2021), https://arxiv.org/abs/2101.06480 3
18. Kimhi, M., Kimhi, S., Zheltonozhskii, E., Litany, O., Baskin, C.: Semi-supervised semantic segmentation via marginal contextual information (2023) 7
19. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023) 10, 11
20. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 10
21. Kroemer, O., Niekum, S., Konidaris, G.D.: A review of robot learning for manipulation: Challenges, representations, and algorithms. J. Mach. Learn. Res. **22**, 30:1–30:82 (2019) 2, 3
22. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 2169–2178 (2006). https://doi.org/10.1109/CVPR.2006.68 6
23. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: European Conference on Computer Vision. pp. 280–296. Springer (2022) 10
24. Li, Y., Zhang, M., Grotz, M., Mo, K., Fox, D.: Stow: Discrete-frame segmentation and tracking of unseen objects for warehouse picking robots. In: Conference on Robot Learning, CoRL 2023 (2023), https://openreview.net/pdf?id=48qUHKUEdBf 2, 3, 4
25. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 936–944 (2017). https://doi.org/10.1109/CVPR.2017.106 5
26. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision (Oct 2017), https://openaccess.thecvf.com/content_iccv_2017/html/Lin_Focal_Loss_for_ICCV_2017_paper.html 5
27. Liu, Y., Chen, X., Abbeel, P.: Self-supervised instance segmentation by grasping. ArXiv **abs/2305.06305** (2023) 2, 4
28. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021) 5, 10, 1
29. Long, Z., Killick, G., McCreadie, R., Camarasa, G.A.: Robollm: Robotic vision tasks grounded on multimodal large language models (2023) 9, 10
30. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019) 2
31. Lu, Y., Chen, Y., Ruozzi, N., Xiang, Y.: Mean shift mask transformer for unseen object instance segmentation. ArXiv **abs/2211.11679** (2022) 11

32. Lu, Y., Chen, Y., Ruozzi, N., Xiang, Y.: Mean shift mask transformer for unseen object instance segmentation (2022). https://doi.org/10.48550/ARXIV.2211.11679 9
33. Milletari, F., Navab, N., Ahmadi, S.A.: 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) 5
34. Mitash, C., Wang, F., Lu, S., Terhuja, V., Garaas, T.W., Polido, F., Nambi, M.: Armbench: An object-centric benchmark dataset for robotic manipulation. 2023 IEEE International Conference on Robotics and Automation (ICRA) pp. 9132–9139 (2023) 1, 3, 9, 10, 2
35. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: 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) 5
36. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: FixMatch: simplifying semi-supervised learning with consistency and confidence. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 596–608. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/hash/06964dce9addb1c5cb5d6e3d9838f733-Abstract.html 1, 3, 7
37. Suchi, M., Patten, T., Vincze, M.: Easylabel: A semi-automatic pixel-wise object annotation tool for creating robotic rgb-d datasets. 2019 International Conference on Robotics and Automation (ICRA) pp. 6678–6684 (2019), https://api.semanticscholar.org/CorpusID:59604455 1, 3, 9, 11, 2
38. Tang, C., Huang, D., Ge, W., Liu, W., Zhang, H.: Graspgpt: Leveraging semantic knowledge from a large language model for task-oriented grasping. arXiv preprint arXiv:2307.13204 (2023) 2
39. Wang, Y., Chen, H., Heng, Q., Hou, W., Fan, Y., Wu, Z., Wang, J., Savvides, M., Shinozaki, T., Raj, B., Schiele, B., Xie, X.: FreeMatch: self-adaptive thresholding for semi-supervised learning. In: International Conference on Learning Representations (2023), https://openreview.net/forum?id=PDrUPTXJI_A 1
40. Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., Wu, L., Zhao, R., Le, X.: Semi-supervised semantic segmentation using unreliable pseudo labels. In: IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR) (2022), https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Semi-Supervised_Semantic_Segmentation_Using_Unreliable_Pseudo-Labels_CVPR_2022_paper.html 7, 8
41. Wen, B., Lian, W., Bekris, K., Schaal, S.: Catgrasp: Learning category-level task-relevant grasping in clutter from simulation. ICRA 2022 (2022) 3
42. Wen, H., Yan, J., Peng, W., Sun, Y.: Transgrasp: Grasp pose estimation of a category of objects by transferring grasps from only one labeled instance. ArXiv **abs/2207.07861** (2022) 3
43. Wu, J., Jiang, Y., Bai, S., Zhang, W., Bai, X.: Seqformer: Sequential transformer for video instance segmentation. In: ECCV (2022) 2, 3, 6, 1
44. Wu, J., Liu, Q., Jiang, Y., Bai, S., Yuille, A., Bai, X.: In defense of online models for video instance segmentation. In: ECCV (2022) 2, 3, 1
45. Wu, Y., Yu, T., Hua, G.: Tracking appearances with occlusions. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. vol. 1, pp. I–I (2003). https://doi.org/10.1109/CVPR.2003.1211433 6

46. Xiang, Y., Xie, C., Mousavian, A., Fox, D.: Learning rgb-d feature embeddings for unseen object instance segmentation. In: Conference on Robot Learning (2020) 11
47. Xie, C., Xiang, Y., Mousavian, A., Fox, D.: Unseen object instance segmentation for robotic environments. IEEE Transactions on Robotics (T-RO) (2021) 3
48. Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q.V.: Unsupervised data augmentation for consistency training. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 6256–6268. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/hash/44feb0096faa8326192570788b38c1d1-Abstract.html 1
49. Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3060–3069 (Oct 2021), https://openaccess.thecvf.com/content/ICCV2021/html/Xu_End-to-End_Semi-Supervised_Object_Detection_With_Soft_Teacher_ICCV_2021_paper.html 8
50. Yang, L., Qi, L., Feng, L., Zhang, W., Shi, Y.: Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In: CVPR (2023) 1
51. Ying, K., Zhong, Q., Mao, W., Wang, Z., Chen, H., Wu, L.Y., Liu, Y., Fan, C., Zhuge, Y., Shen, C.: CTVIS: Consistent Training for Online Video Instance Segmentation (2023) 2, 4
52. Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T.: Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 18408–18419. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper/2021/hash/995693c15f439e3d189b06e89d145dd5-Abstract.html 1, 3
53. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020) 3, 5, 10, 1
54. Ziegler, A., Asano, Y.M.: Self-supervised learning of object parts for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14502–14511 (2022) 3
55. Zini, S., Buzzelli, M., Twardowski, B., van de Weijer, J.: Planckian jitter: enhancing the color quality of self - supervised visual representations. arXiv preprint arXiv: 2202.07993 (2022) 7
56. Zong, Z., Song, G., Liu, Y.: DETRs with collaborative hybrid assignments training. arXiv preprint (Nov 2022), https://arxiv.org/abs/2211.12860 9
57. Zoph, B., Ghiasi, G., Lin, T.Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q.V.: Rethinking pre-training and self-training. ArXiv **abs/2006.06882** (2020) 3, 1

Supplementary Materials for Robot Instance Segmentation with Few Annotations for Grasping

A Technical details

Model The RISE framework begins with an image augmentation step that feeds into a feature extractor followed by an instance segmentation model, and ends at prediction heads for class, bounding box, mask and instance association. We use ResNet-50, ResNet-101 [14] and Swin-L transformer [28] as backbones throughout our experiments, followed by Deformable DETR [53] with 6 encoders and decoders, width of 256 and 300 fixed instance queries, converging on an FPN-like dynamic mask head (as in SeqFormer [43]). In our evaluation, we measure the contribution of the proposed approach to Deformable DETR which serves baseline, and all feature extractors are pretrained on COCO instance segmentation, as is common in Instance segmentation pretraining [57]. The proposed method incorporates a contrastive head (inspired by IDOL [44]) and introduces instance bank, self-supervision branch for non-labeled data, coupled prediction heads for stability (M2B) and label matching strategy during training (MLM). These, in aggregate, allow RISE to outperform both Deformable DETR and SAM, even when these are trained on $\times 10$ more data (1% vs 10%).

Hyperparameters Recall from Sec. 3.1 and Sec. 3.3 that the supervised loss $\mathcal{L}_{\text{s}}$ and unsupervised loss $\mathcal{L}_{\text{u}}$ (Eq. (1), Eq. (5), respectively) are a combination of the class loss $\mathcal{L}_{cls}$, bounding-box loss $\mathcal{L}_{box}$ weighted by λ_1, and the mask loss $\mathcal{L}_{mask}$ weighted by λ_2. We set the loss weights to be $\lambda_1 = 2.0$, $\lambda_2 = 1.0$. The total loss $\mathcal{L}_{\text{total}}$ (in Eq. (9)) combines the supervised loss $\mathcal{L}_{\text{s}}$ or unsupervised loss $\mathcal{L}_{\text{u}}$ (depending on availability of label $\mathbf{y}$), with association loss $\mathcal{L}_{\text{embed}}$ weighted by λ_3. Tab. 6 details an ablation of values of λ_3, showing that the $\mathcal{L}_{\text{embed}}$ contributes to performance, with the best results attained for$\lambda_3 = 0.05$.

Table 6: Ablation of weight λ_3 applied to the sequence association loss $\mathcal{L}_{\text{embed}}$ described in Eq. (4). This evaluation uses 10% of ARMBench labels (90% treated as unlabeled data) and the Swin-L as backbone. A value of $\lambda_3 = 0$ corresponds to a variant that ignores $\mathcal{L}_{embed}$. The best performance is obtained for $\lambda_3 = [0.05, 0.1]$.

λ_3	AP	AP_{50}	AP_{75}
0	74.7	84.9	75.9
0.02	74.4	84.5	75.6
0.05	**74.9**	**85.2**	76.0
0.1	74.5	83.8	**76.2**
0.5	74.3	84.2	75.3

Augmentation Strategy The input images are downsampled and randomly cropped so that the longest side is at most 600 pixels, and so that the shortest side is at least 480 pixels. Recall from Sec. 3.2 that the instance bank contains object cutouts from the labeled portion of the dataset, inspired by [7]. We randomly insert K instances from the instance bank into the image to produce the "before" image x_1 and apply weak augmentations (e.g. slight rotation, translation, brightness etc.). However, we depart from previous approaches by having these K instances distributed according to $Beta(\alpha = 0.5, \beta = 0.5)$, depicted in Fig. 6, making it less likely for synthetically placed instances to occlude actual objects in the image. In addition, we also ensure that the K inserted instances don't overlap with one another beyond 85% since they form ground truth labels during self-supervision. We then generate the "after" image x_2 by randomly adding more instances from the instance bank, or alternatively removing (or transforming) already inserted instances, followed by another round of weak augmentations. The before and after frames serve toward learning through interaction, and we facilitate self-supervised learning by strongly augmenting x_1 to yield x_3 and treat $x_w = x_1$ and $x_s = x_3$ as an input a pair of weakly- and strongly-augmented images. We employ this approach in our evaluation of both ARMBench [34] and OCID [37] without the needing to tune its parameters specific datasets.

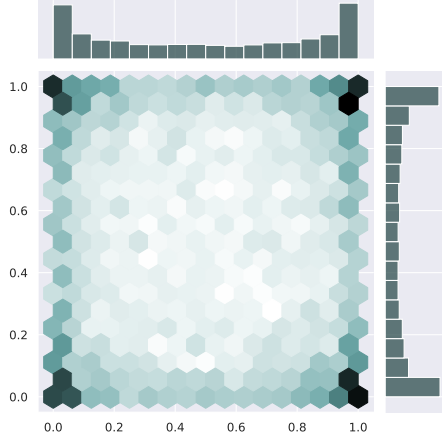


Fig. 6: Two dimensional independent $Beta(\alpha = 0.5, \beta = 0.5)$ distribution representing the spread of instance-bank objects inserted into unlabeled images. The distribution favors placing inserted objects at the periphery of the image, since most images contain most of their information about their center (bright regions denote low probability).

Training. We train our model for $12,000$ iterations, using AdamW [30] optimizer with learning rate of 10^{-4}, and weight decay of 10^{-4} and lr scheduler of StepLR that steps down an order of magnitude after $8,000$ iterations.

B Prediction Matching

The model predicts up to 300 instance labels, boxes and masks which are often far beyond the actual instance count in a given image. In order to compute the loss between valid predictions and ground-truth annotations, we compute the bipartite cost matrix which measures the IoU of each prediction against each ground-truth annotation (either based on box IoU or using the Mask-to-Box method detailed in Sec. 3.3). We then find the fitting assignment for each ground-truth annotation by solving an Optimal Transport (OT) Problem [9]. A similar approach described in Sec. 3.2 serves toward computing $\mathcal{L}_{\text{embed}}$ which requires positive and negative views of an instance. We introduce a method inspired by IDOL [44], where the top-10 prediction matches of each ground-truth annotation are treated as positive views and the rest are considered negative views. The impact of matching is evident in the ablation study in *Tab.* 3 where we use either OT or a more standard approach of using the top 0.7 IoU as positive and bottom 0.3 IoU as negative.

This flow is similarly applied during the self-supervision phase, with the distinction of using pseudo- labels, boxes and masks instead of manually annotated ground truth. Here we also employ Multi-Label Matching (MLM, Sec. 3.3) to allow the model to learn from multiple pseudo-labels predicted from the weak augmentation x_w. The impact of MLM is demonstrated in *Tab.* 3 and inspired by [1], where it further contributes to the framework's performance.

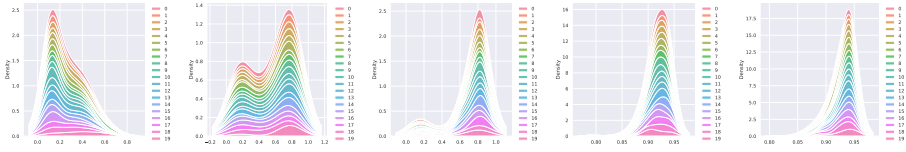
C Thresholds

We use time-dependent thresholds [17, 52], whereby an initial threshold value increases every 1000 training steps. The class and mask thresholds start at $\gamma_t^{\text{cls}} = \gamma_t^{\text{mask}} = 0.85$ and peak at 0.98. For the Cascade approach (Sec. 3.3) which combines a lenient threshold followed by a quantile Q_t described in Eq. (6). We set the initial class and mask thresholds to be $\gamma_t^{\text{cls}} = \gamma_t^{\text{mask}} = 0.5$ and peak at 0.85. The quantile Q_t follows the schedule $p_t = a_0 \cdot (1 - {}^t/_T)$ where t is the training step, T denotes the total number of training steps, and $a_0 = 0.995$ is the quantile base value. Upon ranking the model's predicted instances by their class score, only the top p_t are retained, and the rest are discarded.

D Study of Cascade Threshold

We study the behavior of pseudo-label and pseudo-mask Cascade filter strategy (Eq. (6)). We evaluate the per-instance prediction score of the model using different base values for the quantile Q_t of the Cascade threshold. In Fig. 7, each color band represent 1000 iterations. The figure shows that setting the base value of the quantile too low would allow in more false-negatives as pseudo- labels and masks. Alternatively, setting it too high would discard valuable predictions as they don't meet the ranking requirement of the qunatile. Following this evaluation we set the quantile base value to $a_0 = 0.995$, which leads to the most

balanced behavior of discarding false-positive predictions while allowing through true-positives (even when their score would be considered too low by a standard scheduled threshold).



(a) $a_0 = 0.7$ **(b)** $a_0 = 0.9$ **(c)** $a_0 = 0.95$ **(d)** $a_0 = 0.99$ **(e)** $a_0 = 0.995$

Fig. 7: Confidence density over time using different quantile values for the cascade threshold (Eq. (6)). The x-axis represent the score of all samples, the y-axis the valid instance-count (instance density), and the color band correspond to training iterations in increments of 1000. The cascade threshold applies both a time-dependent threshold (which tightens over time) and a time-dependent quantile Q_t (which loosens over time). The base quantile value a_0 is detailed for each subfigure, showing that the best initial value for the quantile is 0.995.

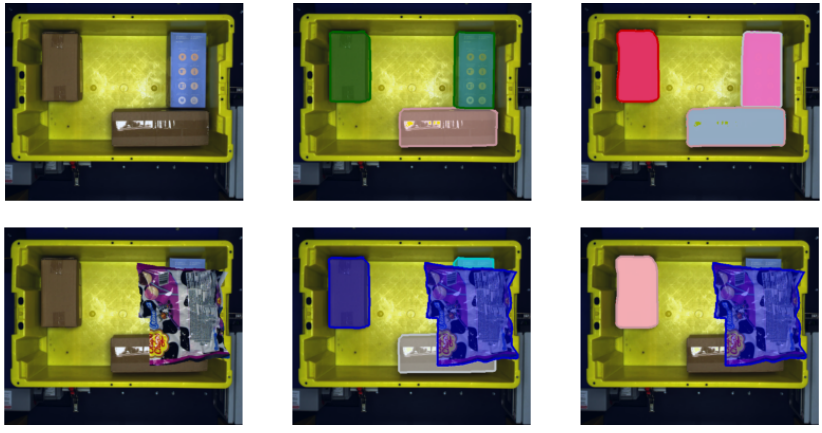


Fig. 8: Failure cases. Mask prediction of a model trained on 1% of annotated data and 99% unlabeled (ResNet-50 backbone). The top row shows that the model accurately predicts the three objects in the tote. The bottom row includes an additional item inserted from the instance-bank, which partially overlaps several objects. Although the model correctly segments the inserted object, it completely misses one occluded object.

E Failure Cases

In both the supervised and self-supervised stages, we randomly draw instance-bank objects and distribute them in the image according to a 2d $Beta(\alpha, \beta)$ distribution (Eq. (2)), and prevent object overlap that exceeds 85% by resampling

from the distribution in case of such overlap. In the supervised phase, we also ensure that inserted objects do not overlap existing (ground-truth) objects by more than 85%, whereas in the self-supervised phase, the $Beta(\alpha, \beta)$ distribution (Eq. (2)) helps reduce the likelihood of inserted memory-bank objects overlapping actual objects in the image (since no ground-truth is available). Despite these precautions, failure cases still occur, particularly at very low annotation rates. Since our method incorporate noisy pseudo–labels in low annotated data regime, we will follow improvements in noisy spatial labels [12] for combating with noise and improve pseudo–labels. Fig. 8 shows how a model trained on 1% of the labeled data (99% treated as unlabeled) accurately predicts the masks of all objects in the "before" image x_1 (and ignores the background). However, in the "after" image x_2 (post-interaction), which contains an additionally inserted object (bottom row), the model fails to produce masks for the occluded cardboard box.