Open-world Instance Segmentation: Top-down Learning with Bottom-up Supervision

Tarun Kalluri[†*] Weiyao Wang[‡] Heng Wang[‡] Manmohan Chandraker[†]
Lorenzo Torresani[‡] Du Tran[‡]

[†]UC San Diego [‡] Meta AI
https://tarun005.github.io/UDOS

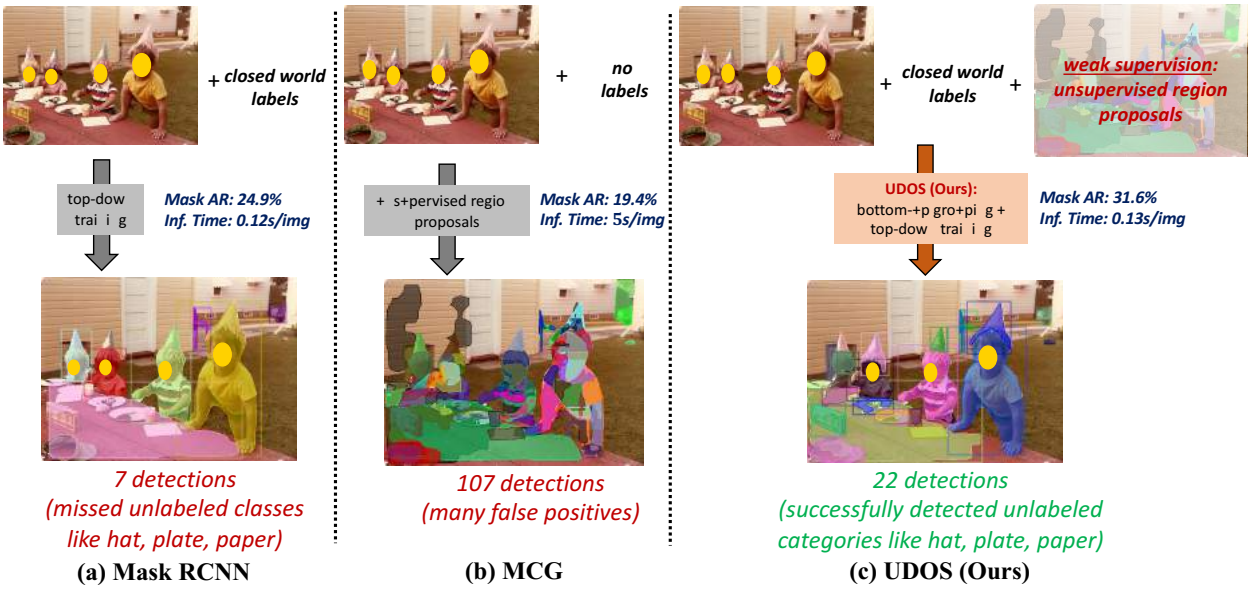


Fig. 1. **Open world segmentation using UDOS**. Image from COCO. **(a)** Mask R-CNN [20], trained on VOC-categories from COCO, fails to detect many unseen categories due to seen-class bias; **(b)** MCG [47] provides diverse proposals, but predicts many over-segmented false-positives with noisy boundaries; **(c)** combining the advantages of (a) and (b) into a joint framework, UDOS efficiently detects unseen classes in open world when trained only using VOC-categories from COCO, while adding negligible inference time overhead.

Abstract

Top-down instance segmentation architectures excel with predefined closed-world taxonomies but exhibit biases and performance degradation in open-world scenarios. In this work, we introduce bottom-**U**p and top-**D**own **O**pen-world **S**egmentation (UDOS), a novel approach that combines classical bottom-up segmentation methods within a top-down learning framework. UDOS leverages a top-down network trained with weak supervision derived from class-agnostic bottom-up segmentation to predict object parts. These part-masks undergo affinity-based grouping and refinement to generate precise instance-level segmentations. UDOS balances the efficiency of top-down architectures with the capacity to handle unseen categories through bottom-up supervision. We validate UDOS on challenging datasets (MS-COCO, LVIS, ADE20k, UVO, and OpenImages), achieving superior performance over state-of-the-art methods in cross-category and cross-dataset transfer tasks. Our code and models will be publicly available.

[*]Work done during TK's internship at Meta.

1. Introduction

Open world instance segmentation [55] is the task of predicting class-agnostic instance masks for all objects within an image. A pivotal challenge therein lies in effectively segmenting novel instances, i.e., instances from categories not in the training taxonomy. This capability assumes paramount importance for ensuring the robust and dependable real-world deployment of instance segmentation models across domains like robotics [57], autonomous driving [11, 43], and embodied AI [50], where novel objects are encountered regularly. While expanding taxonomy during annotation is a potential countermeasure, it presents notable challenges: it necessitates substantial human effort to amass sufficient annotations for each category, and achieving a comprehensive taxonomy encompassing *all* conceivable categories remains an impractical endeavor. Consequently, the emphasis remains on the model's capacity to generalize and proficiently segment novel objects—a more pragmatic approach.

Common instance segmentation frameworks like Mask R-CNN [20] often tightly couple recognition and segmentation [55], making it challenging to accurately segment objects not present in the training data. This issue is particularly pronounced when these frameworks are trained with non-exhaustive annotations like MS-COCO [35], where out-of-taxonomy objects are treated as background, resulting in penalties for predictions made on these objects. In Fig. 1(a), a typical Mask R-CNN model, trained on the 20 VOC classes from the COCO dataset, effectively identifies objects within the training taxonomy such as people and chairs. However, it struggles to detect objects beyond this taxonomy, like hats, paper, and plates.

Conversely, classical bottom-up segmentation methods [18, 47, 53] are class-agnostic and unsupervised by design, making them suitable for open-world scenarios. These methods rely solely on low-level cues such as shape, size, color and texture to generate object masks. However, they often suffer from over-segmentation, lacking a semantic understanding of objectness. In Fig. 1(b), MCG [47] generates over-segmentation of objects with noisy boundaries.

How can we combine the strengths of both paradigms? We answer this question with our novel approach for open-world instance segmentation, termed UDOS (Bottom-Up and Top-Down Open-World Segmentation). UDOS seamlessly integrates the advantages of both top-down and bottom-up methods into a unified and jointly trainable framework. UDOS (Fig. 1c) effectively segments known categories like persons and chairs while demonstrating robust generalization to unseen categories like party hats, paper, glasses, and plates.

UDOS is grounded on two key intuitions: First, we recognize the value of weak supervision from class-agnostic segmentation generated by unsupervised bottom-up methods [18, 47, 53]. This supplementation complements potentially incomplete human annotations, ensuring holistic image segmentation without designating any region as negative. Second, we leverage seen-class supervision to bootstrap objectness, introducing an affinity-based grouping module to merge parts into complete objects and a refinement module to enhance boundary quality. Despite only being trained on seen categories, we observe that both part-level grouping and boundary refinement generalize well to novel categories.

UDOS is the first approach that effectively combines top-down architecture and bottom-up supervision into a unified framework for open-world instance segmentation, and we show its superiority over existing methods through extensive empirical experiments. Our contributions are:

1. We propose UDOS for open-world instance segmentation that effectively combines bottom-up unsupervised grouping with top-down learning in a single jointly trainable framework (Sec. 3).
2. We propose an affinity based grouping strategy (Sec. 3.2) followed by a refinement module (Sec. 3.3) to convert noisy part-segmentations into coherent object segmentations. We show that such grouping generalizes well to unseen objects.
3. UDOS achieves significant improvements over competitive baselines as well as recent open-world instance segmentation methods OLN [29], LDET [49] and GGN [56] on cross-category generalization (VOC to NonVOC) as well as cross-dataset (COCO to UVO, ADE20K and OpenImagesV6) settings (Sec. 4).

2. Related Works

Object detection and instance segmentation. In the past, these tasks relied on handcrafted low-level cues with a bottom-up approach: graph-based grouping [15, 17], graph-based methods [10, 14, 51], deformable parts [16], hierarchical and combinatorial grouping [2, 47] or Selective Search [53]. The rise of deep learning brought about top-down approaches, excelling in various detection and segmentation tasks, including object proposals [33, 46], object detection [48], semantic segmentation [41], instance segmentation [3, 9, 20] and panoptic segmentation [30, 54]. However, this paper addresses a fundamentally different challenge. Instead of assuming a closed-world scenario where training and testing share the same taxonomy, we focus on the open-world, which involves segmenting both in-taxonomy and out-of-taxonomy instances. As in Fig. 1 and Sec. 4.2, top-down methods exhibit a bias towards seen classes and struggle to detect novel objects.

Open world instance segmentation. Open-world vision involves generalizing to unseen objects [7, 46] and is regaining traction in computer vision [22, 28, 29, 39, 40, 55]. We focus on open-world instance segmentation [55],

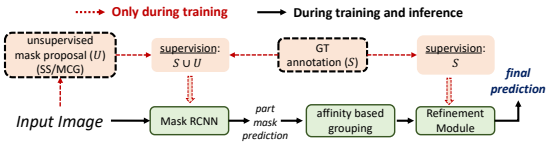


Fig. 2. UDOS overview

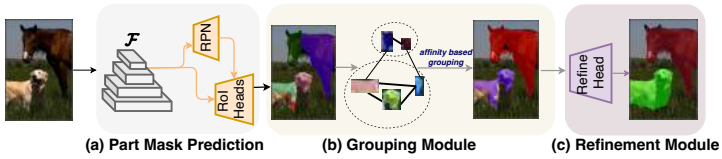


Fig. 3. Proposed UDOS pipeline

(Fig. 2) UDOS overview: Training and inference phases in UDOS. The unsupervised proposal generation is only present during training and not used in inference. **(Fig. 3) Proposed UDOS pipeline:** During training, we first augment the ground truth annotations on seen classes (S) with masks provided by the unsupervised segmentation algorithm (U) and use it to supervise the part mask prediction head in (a) (Sec. 3.1). As these predictions might only correspond to part-segments on unknown classes (*head* of the horse, *body* of the dog), we use an affinity based grouping strategy in (b) that merges part segments corresponding to the same instance (Sec. 3.2 and Fig. 4). We then apply a refinement head in (c) to predict high-quality segmentation for complete instances.

where the goal is to detect and segment objects, even if their categories weren't in the training data. This differs from works [23, 34] that rely on categories with bounding box annotations during training. Prior methods [12, 13, 26, 45] often used additional cues like video, depth, or optical flow. In contrast, UDOS requires no extra annotations and relies on unsupervised proposal generation. Our work is related to [29, 55, 56]. Wang et al. [55] introduces a benchmark, while our paper presents a novel approach. OLN [29] enhances objectness but uses only seen-class annotations in training. UDOS combines top-down training and bottom-up grouping for novel object segmentation. GGN [56] shares a similar approach with UDOS by using bottom-up grouping. However, GGN uses pixel-level pairwise affinities for grouping, while UDOS uses part-level pairwise affinities, with fundamentally different grouping principles. Our results show that UDOS performs well compared to GGN, indicating it could be a complementary method. Finally, UDOS is also complimentary to recent innovations like Segment Anything (SAM) [31], where the potential use of masks generated by SAM as initial segmentations can further improve the quality of open-world detections by UDOS.

Combining bottom-up and top-down. Recent research has revisited bottom-up methods in representation learning [6, 21, 59]. In instance segmentation, bottom-up grouping has improved local segmentation quality using affinity maps [4, 38, 56, 58] or pixel grouping [25, 37, 52]. However, these approaches focus on closed-world taxonomies. Our work combines top-down training and bottom-up grouping for open-world instance segmentation, distinguishing it from prior grouping-based methods [1, 32] that use low-level pixel features. We also address open-world instance segmentation, unlike prior work on 3D part discovery or shape analysis [42].

3. Proposed Method

Problem definition. Given an image $I \in \mathbb{R}^{H \times W \times 3}$, the goal of open world instance segmentation is to segment all object instances in I regardless of their semantic categories, which includes objects that were both seen and unseen during training. Following prior works [29, 46, 56], we adopt class-agnostic learning strategy, in which all annotated classes are mapped to a single foreground class during training and predictions are class-agnostic.

Method overview of UDOS. We visualize the training and inference flows of UDOS in Fig. 2. UDOS consists of part-mask prediction (Fig. 3a), affinity-based grouping (Fig. 3b) and refinement (Fig. 3c). We use class-agnostic Mask R-CNN [20] with FPN [36] as backbone, and we denote the FPN feature map as $\mathcal{F}$.

3.1. Part-Mask Prediction

Generating candidate object regions. We start by creating weak supervision using unsupervised segmentation algorithms (e.g., selective search [53] or MCG [47]) for each image in the training set. These segmentation masks are class-agnostic and cover the entire image, regardless of in-taxonomy or out-of-taxonomy objects. We intentionally favor over-segmentation during proposal generation by tuning the algorithms' hyperparameters (e.g., scale and σ in selective search). It's important to note that this process is a *one-time* effort before training and is not needed during inference (Fig. 3).

Augmenting labels using part-masks. Next, for each training image I, we create a triplet (I, S, U), where $S = \{s_i\}_{i=1}^{N_s}$ represents the set of ground truth box and mask labels for annotated categories, and $U = \{u_i\}_{i=1}^{N_u}$ represents masks generated by the unsupervised segmentation algorithm, offering more extensive but potentially noisy region proposals. We use the augmented masks set $A = S \cup U$ as supervision to train a top-down instance segmentation sys-

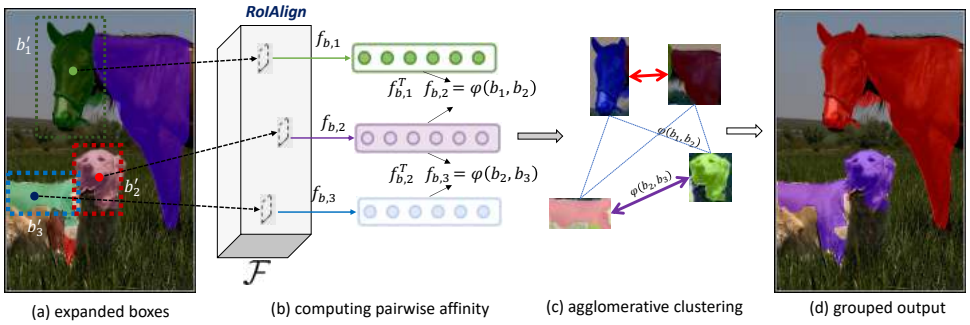


Fig. 4. **Grouping module**. (a) the bounding boxes b_i of the predicted part-masks are expanded to incorporate local context. (b) The features $f_{b,i}$ are extracted using RoIAlign operator on the FPN features $\mathcal{F}$ with the expanded bounding boxes b'_i, and are used to compute pairwise affinity $\phi(b_i, b_j)$ using cosine similarity. (c) A clustering algorithm is used to group parts into whole object instances, as shown in (d). Note that the inaccuracies in the output from grouping module are later corrected by the refinement module.

tem, referred to as the part-mask prediction network. This network may predict only parts of objects in alignment with the provided supervision (output Fig. 3a). To avoid label duplication, we exclude part masks from U that overlap with any ground truth mask in S with an IoU greater than 0.9. In essence, while masks in S provide information for detecting in-taxonomy classes, masks in U assist in segmenting part masks for **all** objects, offering complementary training signals to the network. This strategy offers two key advantages over top-down training with only ground truth masks in S. First, unsupervised region proposals from U account for un-annotated image regions that may contain valid out-of-taxonomy objects, preventing the network from mistakenly labeling them as background. Second, despite masks in U potentially not representing complete objects, they still provide valuable training signals for detecting out-of-taxonomy objects. For example, accurately detecting parts of a dog, such as the head, body, and ears in Fig. 3, proves useful in the final segmentation of the entire dog through our part-mask grouping strategy.

3.2. Grouping Module

To bridge the gap between mid-level part-masks (Sec. 3.1) and complete object instances, we propose an efficient lightweight grouping strategy to merge parts into objects. We compute pairwise affinities between features of the *expanded* parts, and cluster them based on affinities.

Pairwise affinity We denote the predictions made by the network in the first phase by $P = \{p_i\}_{i=1}^{n_p}$, where n_p is the number of predictions, and p_i contains mask (m_i) and box (b_i) predictions made on seen as well as unseen categories. For each bounding box $b_i \in p_i$, we first expand the width and height of the box by a factor $\delta (0<\delta<1)$ to compute a new, larger bounding box b'_i (Fig. 4a).

$$b_i : (x_i, y_i, h_i, w_i) \xrightarrow{\text{expand}} b'_i : (x_i, y_i, (1+\delta)*h_i, (1+\delta)*w_i) \quad (1)$$

where (x_i, y_i) is the center and (h_i, w_i) are the original height and width of box b_i. This inflation allows us to ingest useful context information around the part and the underlying object, better informing the affinities between the part-masks. Next, we compute the ROIAlign features for all the boxes $\{b'_i\}$ from the FPN feature map $\mathcal{F}$ resulting in a d-dim feature for each part-prediction denoted using $\{f_{b,i}\}_{i=1}^{n_p} \in \mathbb{R}^d$. The pairwise affinity between two part-predictions $(p_i, p_j) \in P$ is then computed using the cosine similarity between the corresponding feature maps (Fig. 4b).

$$\phi(p_i, p_j) = \frac{{f_{b,i}}^T \cdot f_{b,j}}{\|f_{b,i}\| \|f_{b,j}\|}; f_{b,i} = \text{RoIAlign}(\mathcal{F}, b'_i) \quad (2)$$

We visualize the parts retrieved using pairwise affinity for few examples in Fig. 6. While Wang et al. [56] has shown strong generalization of pixel pairwise affinities, our novelty lies in showing that the part-mask pairwise affinities generalize better across object categories.

Affinity based grouping We use a clustering algorithm to merge parts based on the soft affinity scores given in Eq. (2). Our clustering objective can be formulated as follows:

$$\max_{G} \sum_{k=1}^{|G|} \sum_{p_i, p_j \in g_k} \phi(p_i, p_j), \quad \text{s.t.} \sum_{k=1}^{|G|} |g_k| = n_p \quad (3)$$

where G is a possible partition of the n_p predictions, $|G|$ denotes the total number of partitions and k^{th} partition in G is denoted by g_k ($1 \leq k \leq |G|$). In other words, given a set of elements along with their pairwise affinity scores, our clustering algorithm produces a partition of the elements that maximizes the average affinities within each partition. We use an off-the-shelf agglomerative clustering algorithm from Bansal et al. [5] provided by `scikit-learn` [44]. It is parameter-free, lightweight, and fast, incurring minimum time and memory overhead while clustering hundreds of

part-masks in each iteration. As shown in Sec. 4.4 our final framework adds negligible inference time overhead to the backbone. We merge all the part masks (and boxes) within each partition group g_k to form more complete masks, representing whole objects (Fig. 3b). Since the original predictions in P might also represent whole objects on seen classes, we combine the originally detected masks as well as the grouped masks into our output at this stage.

3.3. Refinement Module

To address potential blurriness in the grouped masks due to noisy initial segmentation, we incorporate a refinement module. Designed similar to the RoIHeads in Mask R-CNN, this module takes the predictions generated after the grouping stage as inputs (Fig. 3c). We train the refinement head exclusively on annotated ground truth instances from S to introduce the concept of object boundaries into the predictions (only available in the annotated masks). We found that this boundary refinement also generalizes well to unseen categories. We jointly train the backbone and refinement heads in a single stage, using losses from the part-mask prediction and refinement modules.

Objectness ranking Following [29], we add box and mask IoU branches to our RoIHead in part-mask predictions as well as refinement heads to compute the localization quality. IoU metrics are shown to improve objectness prediction [24] and avoid over-fitting to seen instances [29] when trained with non-exhaustive annotations. We use box and mask IoU heads with two fc-layers of 256-dim each followed by a linear layer to predict the IoU score, trained using an L1 loss for IoU regression.

Inference (Fig. 2). We first predict part masks, followed by the affinity based grouping to hierarchically merge them into complete objects. We then pass these detections through the refinement layer to obtain the final predictions. We rank the predictions using the geometric mean of their predicted classification score c, box IoU b and mask IoU m from the refinement head as $s = \sqrt[3]{c * b * m}$.

4. Experiments

Datasets and evaluations. We demonstrate the effectiveness of UDOS for open-world instance segmentation under *cross-category* generalization within the same dataset, as well as *cross-dataset* generalization across datasets with different taxonomies (Tab. 1). We use the MS-COCO [35] for cross-category generalization and train the model using 20 categories from VOC and test on the 60 remaining unseen nonVOC classes following prior work [29, 49, 56]. For cross-dataset generalization, we train on complete COCO dataset and directly test on validation splits of UVO [55], ADE20k [60] and OpenImagesV6 [8] datasets without any fine-tuning. We also test large-taxonomy scenario by training on a subset of 1123 categories from LVIS [19] and test on COCO. Both UVO and ADE20k datasets provide exhaustive annotations in every frame, which is ideal to evaluate open world models, while OpenImagesV6 with 350 categories allows to test our open world segmentation approach on large scale datasets.

Cross-category setting			
Train On	Test On	# Seen classes	# Unseen classes
VOC	Non-VOC	20	60
LVIS	COCO	1123	80
Cross-dataset setting			
Train On	Test On	# Seen classes	# Unseen classes
COCO	UVO	80	open
	ADE20k		70
	OpenImagesV6		270

Tab. 1. **Evaluation settings**. Seen and unseen categories used in our evaluation.

Implementation details. We use Mask R-CNN model [20] with a ResNet-50-FPN [36] as our backbone. We train UDOS using SGD for 10 epochs with an initial learning rate of 0.02 on 8 GPUs. We use selective search [53] to generate unsupervised masks for images in COCO dataset. Note that the mask proposals are required *only during training*, and *not* during inference (Fig. 2). We follow prior works in open-world instance segmentation [29, 49, 56] and use average recall (AR) (between IoU thresholds of 0.5 to 1.0) as the evaluation metric. Since open world models generally detect many more objects in a scene than closed world models (see Fig. 5) and many datasets do not have exhaustive annotation, we use AR^{100} and AR^{300} as the evaluation metrics on both box (AR_B) and mask (AR_M) to avoid penalizing predictions of valid, yet unannotated, objects.

4.1. Baselines

(i) **Image-computable masks**: We use masks generated by **MCG** [47] and Selective Search [53] (**SS**), which are class-agnostic, learning-free proposal generation methods relying on low-level cues. (ii) **Mask-RCNN** [20] denotes Mask R-CNN training in class-agnostic fashion only on the seen classes, (iii) **Mask R-CNN**$_{SS}$ indicates Mask R-CNN trained using selective search proposals as the supervision instead of the ground truth annotations, and (iv) **Mask R-CNN**$_{SC}$ denotes Mask R-CNN trained with BoxIoU and MaskIoU scoring to rank the proposals instead of the classification score.

We also compare with state of the art open-world instance segmentation algorithms OLN [29], LDET [49] and GGN [56]. For fair comparison with UDOS, we use the result from GGN [56] *without* the OLN backbone.

VOC→NonVOC	AR_B^{100}	AR_B^{300}	AR_M^{100}	AR_M^{300}
Bottom-up*(No Training)*				
SS	14.3	24.7	6.7	12.9
MCG	23.6	30.8	19.4	25.2
Top-down*(Class-agnostic Training)*				
MaskRCNN	25.1	30.8	20.5	25.1
Mask R-CNN$_{SS}$	24.1	24.9	20.9	21.7
Mask R-CNN$_{SC}$	25.6	33.1	24.9	28
Open-World Methods				
OLN	32.5	37.4	26.9	30.4
LDET	30.9	38.0	26.7	32.5
GGN	31.6	39.5	28.7	35.5
UDOS	**33.5**	**41.6**	**31.6**	**35.6**

Tab. 2. **Cross-category generalization evaluation on COCO.** Train on 20 VOC categories and test on 60 NonVOC categories. UDOS outperforms many competitive baselines as well as the current SOTA GGN on the VOC→NonVOC setting.

4.2. UDOS outperforms baselines on cross-category generalization

Existing methods relying on bottom-up grouping with techniques like SS or MCG, or top-down architectures like Mask R-CNN trained on annotations for seen classes, struggle to effectively detect and segment instances from unseen classes. In contrast, UDOS overcomes these limitations (Tab. 2) by harnessing both ground truth annotations for known classes and unsupervised bottom-up masks for unknown classes, resulting in a significant improvement over all baseline methods, underscoring the effectiveness of our approach. From Tab. 2, UDOS achieves Box Average Recall (AR) of 33.5% and Mask AR of 31.6% setting a new state-of-the-art in cross-category open-world instance segmentation, outperforming the current state-of-the-art methods like GGN in both box and mask AR.

Expanding training datasets to encompass larger taxonomies is a potential strategy for addressing the challenge of novel categories at test-time. However, our experiments, detailed in Tab. 3, reveal that this approach still falls short of achieving robust generalization to unseen categories. We leveraged the LVIS dataset [19], which includes annotations for 1203 categories. During training, we excluded annotations with an IoU overlap greater than 0.5 with COCO masks, resulting in 79.5k instance masks from LVIS. When evaluated on COCO validation images, UDOS achieved 33.2% AR_B^{100} and 26.3% AR_M^{100}, markedly outperforming the baseline methods. This underscores the effectiveness of UDOS in even handling datasets with large category vocabularies.

4.3. UDOS sets new SOTA on cross-dataset generalization

To provide a more realistic assessment of our model's open-world capabilities, we evaluated its performance on real-world target datasets like UVO, ADE20k, and OpenImages. These datasets contain a wide range of objects, including those not covered by COCO categories. It's worth noting that we refrained from fine-tuning our model on the target datasets or using any unlabeled target data during training. Results are in Tab. 4, and summarized below.

LVIS→COCO	AR_B^{100}	AR_B^{300}	AR_M^{100}	AR_M^{300}
MaskRCNN	23.8	29.4	18.5	22.0
Mask R-CNN$_{SC}$	21.3	27.9	17.9	24.2
OLN [29]	28.5	38.1	23.4	27.9
UDOS	**33.2**	**42.2**	**26.3**	**32.2**

Tab. 3. **Cross-category generalization evaluation with large taxonomy**. All models are trained on 1123 categories from LVIS (excluding COCO categories), and evaluated on COCO 80 categories. UDOS outperforms OLN [29] by 4.7% and 2.9% on box and mask AR^{100}.

COCO to UVO. Since UDOS is designed to handle novel classes, it achieves much better performance than other baselines on the challenging UVO dataset that contains exhaustive annotations for every objects. UDOS clearly outperforms baseline approaches like Mask R-CNN (+5% AR_B^{100}). We also outperform OLN, LDET and GGN, setting the new state-of-the-art on the UVO benchmark.

COCO to ADE20K ADE20k [60] is a scene parsing benchmark consisting of annotations for both *stuff* (road, sky, floor etc.) and discrete *thing* classes. We regard each annotation mask as a separate semantic entity and compute the average recall (AR) on both in-taxonomy and out-of-taxonomy objects to evaluate the ability of trained models to detect thing classes and group stuff classes in images. From Tab. 4, we observe that UDOS achieves box AR100 of 27.2% and mask AR100 of 23.0%, higher than all the baselines and other competing methods.

COCO to OpenImagesV6 Again, UDOS consistently outperform all baselines as well as open-world methods like OLN and GGN by significant margins on the OpenImagesV6 dataset [8]. We achieve AR_B^{100} of 71.6%, which is better than the strongest baseline Mask R-CNN by 14.5% and current state-of-the-art GGN by 7.1%. Likewise, AR_M^{100} of 66.2% obtained by UDOS is 4.8% higher than GGN, setting new state of the art.

4.4. Ablations

We use the VOC to NonVOC cross-category generalization on COCO dataset for the ablations.

Refinement and grouping modules We show in Tab. 5a that without the proposed grouping and refinement modules, maskAR drops to 11.8% from 31.6%, as the masks are noisy and only correspond to parts of instances. Using a refinement module after grouping leads to more refined masks further improving the performance.

Choice of proposal ranking We show the importance of using BoxIoU and MaskIoU scoring functions in Tab. 5b,

	COCO→UVO				COCO→ADE20K				COCO→OpenImages			
	AR_B^{100}	AR_B^{300}	AR_M^{100}	AR_M^{300}	AR_B^{100}	AR_B^{300}	AR_M^{100}	AR_M^{300}	AR_B^{100}	AR_B^{300}	AR_M^{100}	AR_M^{300}
MaskRCNN	47.7	50.7	41.1	43.6	18.6	24.2	15.5	20.0	57.1	59.1	55.6	57.7
Mask R-CNN$_{SS}$	26.8	31.5	25.1	31.1	18.2	25.0	17	21.6	34.0	42.7	33.1	38.8
Mask R-CNN$_{SC}$	42.0	50.8	40.7	44.1	19.1	25.6	18.0	22.0	54.1	59.1	54.2	57.4
OLN	50.3	57.1	41.4	44.7	24.7	32.1	20.4	27.2	60.1	64.1	60.0	63.5
LDET	52.8	58.7	43.1	47.2	22.9	29.8	19.0	24.1	59.6	63.0	58.4	61.4
GGN	52.8	58.7	43.4	47.5	25.3	32.7	21.0	26.8	64.5	67.9	61.4	64.3
UDOS	**53.6**	**61.0**	**43.8**	**49.2**	**27.2**	**36.0**	**23.0**	**30.2**	**71.6**	**74.6**	**66.2**	**68.7**

Tab. 4. **Cross-dataset generalization evaluation for open world instance segmentation**. All models are trained on 80 COCO categories and evaluated on UVO (left), ADE20K (middle), OpenImages (right) as is without any fine-tuning.

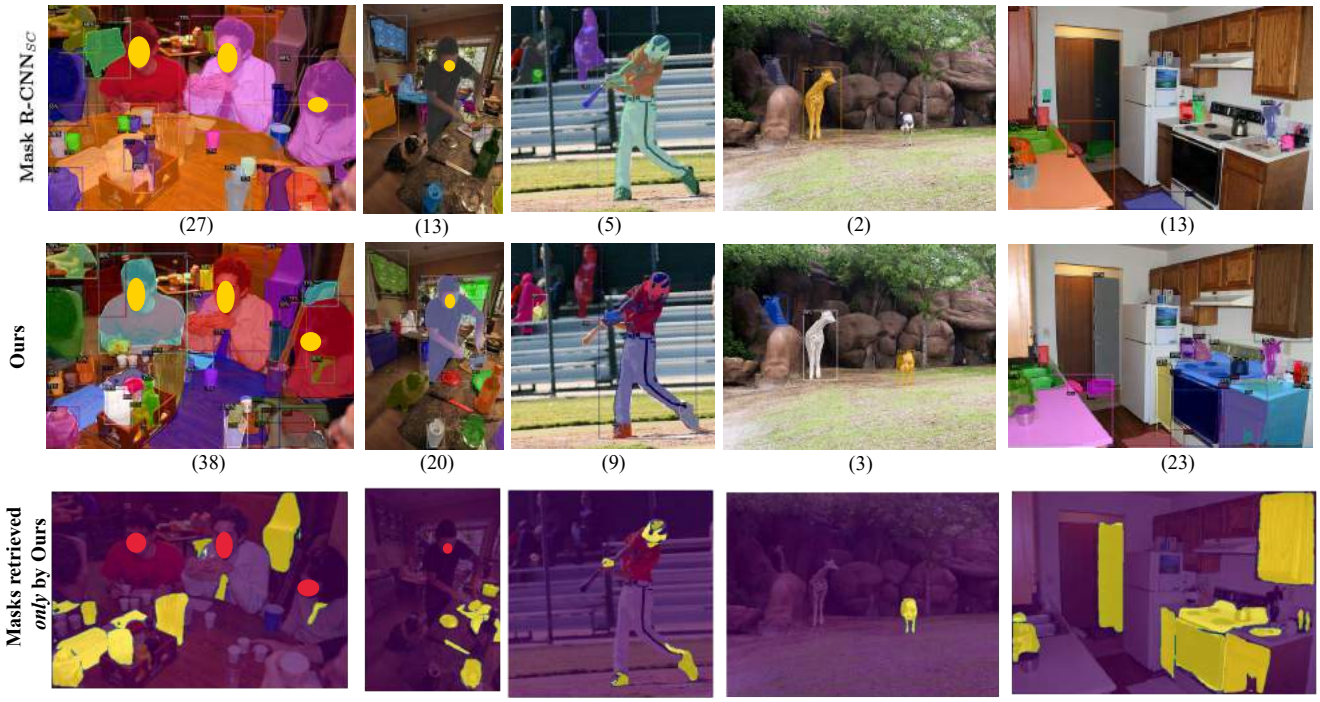


Fig. 5. **Visualization of segmentations for model trained only on VOC classes** from COCO dataset. The top row shows result using using Mask-RCNN$_{SC}$, second row shows output using UDOS and the third row shows some predictions made only by UDOS and missed by Mask-RCNN$_{SC}$. We also show the number of detections made by the network below each image. Starting from left most image, many classes like {jug, tissue papers, tie, eyeglasses}, {knife, cutting board, vegetables, glass}, {shoes, helmet, gloves}, {ostrich} and {dishwasher, faucet} among others which are not part of VOC-classes are missed by standard Mask-RCNN training, but detected using UDOS. More visualizations are provided in the supplementary.

where significant drops in AR100 are observed without the use of both the scoring functions, validating the observations in prior works [29] that scoring module prevents overfitting and improves open world learning.

Influence of δ Intuitively, a small value of delta (part-mask expansion factor, Eq. (1)) would not capture sufficient context around the region for extracting similarity while a very high value of δ would induce noisy features from different neighboring objects. In Tab. 5d, we show that a value of 0.1 achieves an optimum trade-off, so we use the same value of $\delta = 0.1$ in all our experiments.

Choice of proposal generation From Tab. 5c, we show that a naive segmentation of image using uniform grids by extracting 64×64 patches from the image expectedly performs worse, as these part masks do not semantically correspond to object parts. We also use super-pixels generated from SSN [27], but found that bottom-up supervision generated from image-based segmentation algorithms like SS or MCG lead to much better accuracies.

Visualizations of affinity maps In Fig. 6, we present 3-nearest part masks retrieved for a given query mask using their affinity (Eq. (2)) and the grouped outputs. We observe

Group	Refine	AR_B^{100}	AR_M^{100}
✗	✗	25.4	11.8
✓	✗	32.6	30.7
✓	✓	33.5	31.6

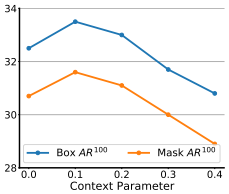
(a)

BoxIoU	MaskIoU	AR_B^{100}	AR_M^{100}
✗	✗	29.0	24.3
✓	✗	32.7	28.9
✗	✓	32.9	29.2
✓	✓	**33.5**	**31.6**

(b)

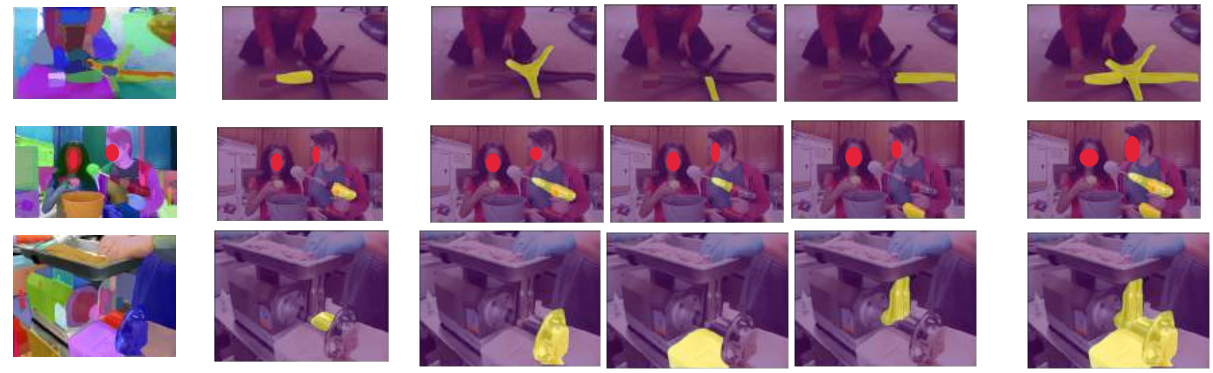
Segmentation	AR_B^{100}	AR_M^{100}
Uniform Grid	9.9	9.2
SSN	19.4	18.7
Sel. Search	**33.5**	**31.6**
MCG	32.4	29.4

(c)



(d)

Tab. 5. **Ablation results**. Effect of ((a)) grouping and refinement modules, ((b)), boxIoU and maskIou losses during training, ((c)) segmentation algorithm and ((d)) context dilation parameter δ on the VOC→NonVOC performance.



(a) part prediction **(b) query mask** **(c) retrieved part masks with highest affinity** **(d) grouped output**

Fig. 6. **Visualization of pairwise affinity maps and grouped predictions.** Given a part mask as a query, we show the 3 nearest part masks of the query using our pairwise affinity. The images are taken from UVO dataset, and the affinity is computed using UDOS model trained on COCO. Our affinity-based grouping module correctly groups parts into whole instances even with unseen objects. The last row visualizes a failure case where the model retrieves a part mask from a neighboring instance.

that different part masks of the same entity are often retrieved with high affinity, using our grouping module.

Inference time comparison Our grouping module is lightweight and adds negligible run-time overhead. Specifically, at 100 output proposals, MaskRCNN [20] and GGN [56] take 0.09s/im, MaskRCNN$_{SC}$ and OLN [29] take 0.12s/im while UDOS takes 0.13s/im (+0.01s/im) with stronger performance. Generating part-masks using selective search for the complete COCO [35] dataset takes around 1.5 days on a 16-core CPU, but we reiterate that the part-masks only need to be generated once before training and are not needed during testing/deployment (Fig. 2). We will publicly release the part-masks on COCO dataset along with the code and trained models.

5. Discussion

In this paper, we conduct an investigation to understand what types of top-down learning generalize well in the context of open-world segmentation. Our observation revealed that learning from known classes to group part-level segmentations and learning to refine coarse boundaries are effective for generalizing to new categories. This allowed us to leverage classical bottom-up segmentation algorithms that provide class-agnostic yet coarse and over-segmented part masks within a top-down learning framework. We introduced UDOS, a novel approach that integrates top-down and bottom-up strategies into a unified framework. Our grouping and refinement modules efficiently convert part mask predictions into complete instance masks for both familiar and unfamiliar objects, setting UDOS apart from previous closed-world and open-world segmentation methods. Extensive experiments demonstrated the significant performance improvements achieved by UDOS across five challenging datasets, including COCO, LVIS, ADE20k, UVO, and OpenImages. We believe that UDOS can serve as a versatile module for enhancing the performance of downstream tasks, such as robotics and embodied AI, in handling unfamiliar objects in open-world scenarios.

Limitations. UDOS faces challenges in scenarios with densely clustered objects of similar appearance. A more robust, learnable grouping method, possibly trained with hard negatives, could enhance performance in such complex situations, suggesting a direction for future research. Furthermore, incorporating recent innovations like Segment Anything (SAM) [31] to improve initial segmentations in UDOS is also en exiciting future directions to follow.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4981–4990, 2018. 3

[2] Pablo Arbelaez. Boundary extraction in natural images using ultrametric contour maps. In *CVPR Workshops*, 2006. 2

[3] Anurag Arnab and Philip HS Torr. Pixelwise instance segmentation with a dynamically instantiated network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 441–450, 2017. 2

[4] Alberto Bailoni, Constantin Pape, Steffen Wolf, Thorsten Beier, Anna Kreshuk, and Fred A Hamprecht. A generalized framework for agglomerative clustering of signed graphs applied to instance segmentation. *arXiv preprint arXiv:1906.11713*, 2019. 3

[5] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine learning*, 56(1):89–113, 2004. 4

[6] Amir Bar, Xin Wang, Vadim Kantorov, Colorado J Reed, Roei Herzig, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Detreg: Unsupervised pretraining with region priors for object detection. *arXiv preprint arXiv:2106.04550*, 2021. 3

[7] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1893–1902, 2015. 2

[8] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11700–11709, 2019. 5, 6

[9] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019. 2

[10] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239, 2001. 2

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 2

[12] Achal Dave, Pavel Tokmakov, and Deva Ramanan. Towards segmenting anything that moves. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[13] Yuming Du, Yang Xiao, and Vincent Lepetit. Learning to better segment objects from unseen classes with unlabeled videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3375–3384, 2021. 3

[14] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2008. 2

[15] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004. 2

[16] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009. 2

[17] M. Grundmann, V. Kwatra, M. Han, and I. Essa. Efficient hierarchical graph-based video segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2141–2148, 2010. 2

[18] Matthias Grundmann, Vivek Kwatra, Mei Han, and Irfan Essa. Efficient hierarchical graph-based video segmentation. In *2010 ieee computer society conference on computer vision and pattern recognition*, pages 2141–2148. IEEE, 2010. 2

[19] 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*, pages 5356–5364, 2019. 5, 6

[20] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2, 3, 5, 8

[21] Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10086–10096, 2021. 3

[22] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10960, 2020. 2

[23] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4233–4241, 2018. 3

[24] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6409–6418, 2019. 5

[25] Jyh-Jing Hwang, Stella X Yu, Jianbo Shi, Maxwell D Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7334–7344, 2019. 3

[26] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2117–2126. IEEE, 2017. 3

[27] Varun Jampani, Deqing Sun, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Superpixel sampling networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 352–368, 2018. 7

[28] K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, 2021. 2

[29] 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*, 2022. 2, 3, 5, 6, 7, 8

[30] Alexander Kirillov, Kaiming He, Ross B. Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019. 2

[31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3, 8

[32] Shu Kong and Charless C Fowlkes. Recurrent pixel embedding for instance grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9018–9028, 2018. 3

[33] W. Kuo, B. Hariharan, and J. Malik. Deepbox: Learning objectness with convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2479–2487, Los Alamitos, CA, USA, 2015. IEEE Computer Society. 2

[34] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. Shapemask: Learning to segment novel objects by refining shape priors. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 9207–9216, 2019. 3

[35] 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*, pages 740–755. Springer, 2014. 2, 5, 8

[36] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3, 5

[37] Shu Liu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. Sgn: Sequential grouping networks for instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3496–3504, 2017. 3

[38] Yiding Liu, Siyu Yang, Bin Li, Wengang Zhou, Jizheng Xu, Houqiang Li, and Yan Lu. Affinity derivation and graph merge for instance segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 686–703, 2018. 3

[39] Yang Liu, Idil Esen Zulfikar, Jonathon Luiten, Achal Dave, Aljosa Osep, Deva Ramanan, Bastian Leibe, and Laura Leal-Taixé. Opening up open-world tracking. *CoRR*, abs/2104.11221, 2021. 2

[40] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 2

[41] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[42] Tiange Luo, Kaichun Mo, Zhiao Huang, Jiarui Xu, Siyu Hu, Liwei Wang, and Hao Su. Learning to group: A bottom-up framework for 3d part discovery in unseen categories. *arXiv preprint arXiv:2002.06478*, 2020. 3

[43] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 2

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 4

[45] Trung Pham, Thanh-Toan Do, Gustavo Carneiro, Ian Reid, et al. Bayesian semantic instance segmentation in open set world. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–18, 2018. 3

[46] Pedro O Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *Advances in neural information processing systems*, 2015. 2, 3

[47] Jordi Pont-Tuset, Pablo Arbelaez, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):128–140, 2016. 1, 2, 3, 5

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99. Curran Associates, Inc., 2015. 2

[49] Kuniaki Saito, Ping Hu, Trevor Darrell, and Kate Saenko. Learning to detect every thing in an open world. *arXiv preprint arXiv:2112.01698*, 2022. 2, 5

[50] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019. 2

[51] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22 (8):888–905, 2000. 2

[52] Mennatullah Siam, Alex Kendall, and Martin Jagersand. Video class agnostic segmentation benchmark for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2825–2834, 2021. 3

[53] 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. 2, 3, 5

[54] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan L. Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic

segmentation with mask transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 2021. 2

[55] 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*, pages 10776–10785, 2021. 2, 3, 5

[56] 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*, pages 4422–4432, 2022. 2, 3, 4, 5, 8

[57] Christopher Xie, Yu Xiang, Arsalan Mousavian, and Dieter Fox. Unseen object instance segmentation for robotic environments. *IEEE Transactions on Robotics*, 37(5):1343–1359, 2021. 2

[58] Xingqian Xu, Mang Tik Chiu, Thomas S Huang, and Honghui Shi. Deep affinity net: Instance segmentation via affinity. *arXiv preprint arXiv:2003.06849*, 2020. 3

[59] Xiao Zhang and Michael Maire. Self-supervised visual representation learning from hierarchical grouping. *Advances in Neural Information Processing Systems*, 33:16579–16590, 2020. 3

[60] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 5, 6