A Weakly Supervised Amodal Segmenter with Boundary Uncertainty Estimation

Khoi Nguyen and Sinisa Todorovic
Oregon State University
Corvallis, OR 97330, USA
{nguyenkh,sinisa}@oregonstate.edu

Abstract

This paper addresses weakly supervised amodal instance segmentation, where the goal is to segment both visible and occluded (amodal) object parts, while training provides only ground-truth visible (modal) segmentations. Following prior work, we use data manipulation to generate occlusions in training images and thus train a segmenter to predict amodal segmentations of the manipulated data. The resulting predictions on training images are taken as the pseudo-ground truth for the standard training of Mask-RCNN, which we use for amodal instance segmentation of test images. For generating the pseudo-ground truth, we specify a new Amodal Segmenter based on Boundary Uncertainty estimation (ASBU) and make two contributions. First, while prior work uses the occluder's mask, our ASBU uses the occlusion boundary as input. Second, ASBU estimates an uncertainty map of the prediction. The estimated uncertainty regularizes learning such that lower segmentation loss is incurred on regions with high uncertainty. ASBU achieves significant performance improvement relative to the state of the art on the COCOA and KINS datasets in three tasks: amodal instance segmentation, amodal completion, and ordering recovery.

1. Introduction

This paper addresses weakly supervised amodal instance segmentation (WAIS). Our goal is to segment both visible and occluded (amodal) parts of object instances in images. The weak supervision in training provides only ground-truth visible (modal) instance segmentations. Important applications of amodal segmentation include autonomous driving and robot path planning, where identifying the whole spatial extents of partially occluded objects is critical. Considering this problem under weak supervision is also important because human annotators often cannot provide reliable ground truth. For example, different annotators are likely to have very different and sometimes poor guesses of occluded object parts.

There is scant prior work on WAIS. Following recent PCNet [36], our training consists of two stages. First, we use data augmentation to train a common image segmenter – UNet [32] – on manipulated training images to predict their amodal segmentations. As input to UNet, we use the available ground-truth modal segmentation and information about where the data augmentation generated the occlusion in the training image. In the second training stage, UNet's amodal segmentations are taken as a pseudo-ground truth for learning a standard instance segmenter – Mask-RCNN [13], as in [36]. On test images with occlusions, Mask-RCNN trained on the pseudo-ground truth is expected to output correct amodal instance segmentation.

Our contributions are aimed at advancing the first training stage, and include: (1) a new way to exploit the weak supervision for training of UNet; and (2) enabling UNet to estimate uncertainty of the predicted amodal segmentation, and enforcing the training of UNet to explicitly minimize this uncertainty.

Our first contribution is motivated by the following limitation of PCNet [36]. For manipulating training images, as illustrated in Fig. 1, PCNet randomly places an *occluder* object onto an *occludee* object based on their ground-truth modal segmentation masks, and in this way artificially generates the occluded mask of the occludee. Then, as three inputs to UNet, PCNet uses the manipulated training image, the occluded mask of the occludee, and the occluder's mask. However, using the occluder's mask as input to UNet puts the restrictive constraint that the occluder itself cannot be occluded by another object. To address this limitation, PCNet estimates an object ordering graph in the image, and for the occluder selects a union of all objects estimated as closer to the camera than the occludee (i.e., a union of multiple occluders). Our novelty is in replacing the occluder's mask with the occlusion boundary in the input to UNet, as depicted in Fig. 1. Thus, our occluders are allowed to be themselves partially occluded by some other objects. This reduces complexity of estimating the pseudo-ground-truth amodal segmentation, as we do not need to estimate the object ordering graph.

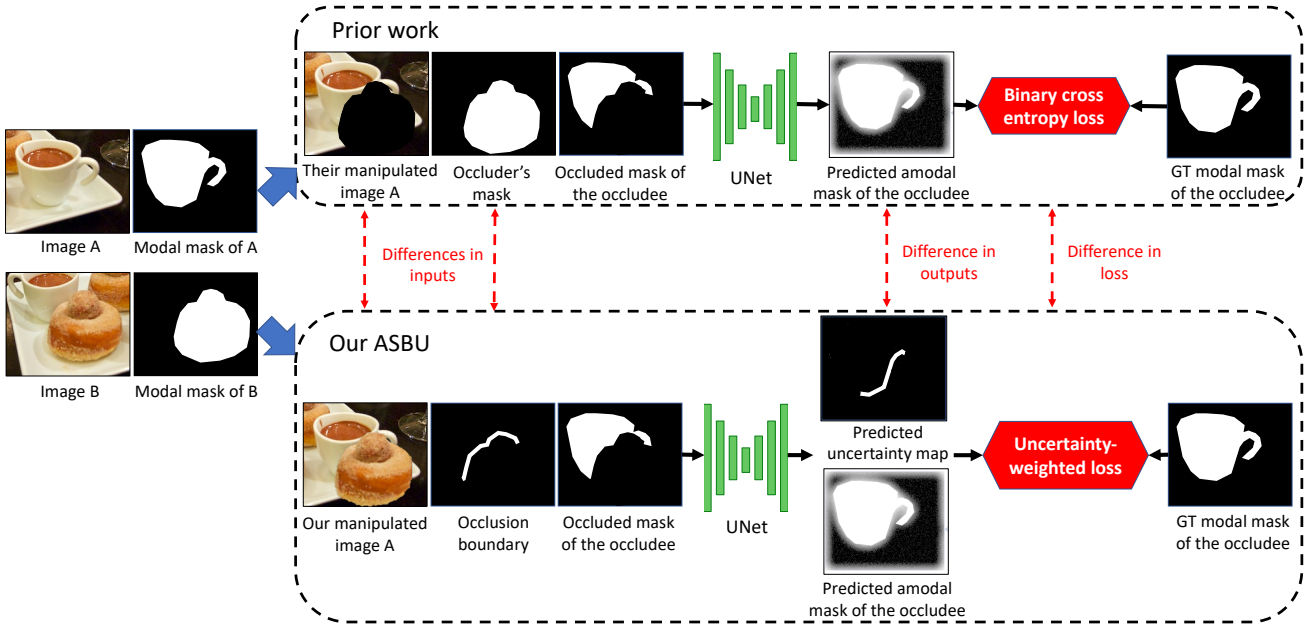

Figure 1. (Top) A recent approach to WAIS [36] where UNet [32] is trained with binary cross-entropy loss to predict amodal segmentation from the manipulated training image, the occluder's mask, and the occluded mask of the occludee. (Bottom) Our approach, called ASBU, differs [36] in terms of input, output, and loss. For input, instead of the occluder's mask, we use the occlusion boundary, and instead of "zeros" we use matting to superimpose the occluder onto the image. This means that we remove the information about the occluder's spatial extent from the input to UNet. For output, along with amodal segmentation, we additionally estimate uncertainty of prediction. For loss, we use uncertainty to appropriately weight loss. GT stands for ground truth.

Our second contribution is aimed at accounting for shape priors. As shown in Fig. 4, we *implicitly* capture a "shape prior" through learning to estimate an uncertainty map for the predicted amodal segmentation. In our experiments, we observe that the estimated uncertainty typically takes low (high) values over regions far away (close) to the occlusion boundary. This suggests that our uncertainty map is capable of representing a spatial distribution of object shapes, and hence can be used for regularizing our learning. Our regularization uses the estimated uncertainty map to appropriately modulate a difference between the predicted amodal segmentation and the original ground-truth mask of the occludee (before the occlusion), such that lower loss is incurred on regions with high uncertainty.

Our two contributions are incorporated in the new Amodal Segmenter with Boundary Uncertainty estimation (ASBU). ASBU is evaluated on the COCOA [37] and KINS [31] datasets on three tasks: amodal instance segmentation, amodal completion, and ordering recovery. ASBU significantly outperforms the state of the art in all three tasks.

In the following, Sec. 2 reviews previous work; Sec. 3 specifies ASBU; Sec. 4 formalizes our uncertainty estimation and uncertainty weighted loss; Sec. 5 presents our implementation details and experimental results; and Sec. 6 concludes the paper.

2. Related Work

This section reviews closely related work.

Instance Segmentation is aimed at labeling pixels with object instance labels, and can be addressed with *bounding-box-based* and *bounding-box-free* methods. In the former [13, 24, 5, 29, 2], for every detected bounding box, a foreground object is segmented. In the latter [20, 22, 3, 10], first, a semantic segmentation is obtained, and then pixels of the same semantic class are clustered into instances based on visual cues such as object center or inner sign distance function. All of these approaches segment only visible object parts and thus are not suitable for our problem.

Amodal Instance Segmentation infers visible and occluded object parts, under full supervision in training. For the ground truth, prior work uses amodal segmentation of either real images [23, 37, 31, 11] or synthetic data [9, 15, 18]. However, the existing quality of synthetic data introduces a domain gap between training on synthetic images and testing on real images, resulting in a considerable performance difference between the two domains.

Amodal Instance Completion differs from amodal instance segmentation since the goal is to complete occluded parts of an object given its modal mask, whereas in amodal instance segmentation the modal mask is not provided. Prior work typically uses the Gestalt principles and makes

certain assumptions about shape convexity and length. For example, amodal instance completion has been addressed by using Euler spiral, cubic Bezier curves, shape primitives, and variational auto-encoder in [19, 25, 34, 28]. Our first stage of training for generating the amodal pseudo-ground truth is based on amodal instance completion. We evaluate ASBU on the task of amodal instance completion.

WAIS provides access only to modal-mask annotations in training. Recent work [36] begins by converting modal-mask annotations of training images into pseudo amodal masks in a self-supervised manner, as illustrated in Fig. 1. However, in a complex scene, the occluder can also be occluded by another object, as shown in Fig. 2, which requires [36] to construct an object ordering graph. This increases complexity of their first stage of training of UNet, and is not even suitable for addressing cases of entangled partial occlusions when the occluder-occludee relationship of a pair of objects is not unique, as illustrated in Fig. 3. We overcome this limitation by replacing the occluder's mask with the occlusion boundary for our input. Unlike [36], we effectively remove from our input any information about the occluder's spatial extent. Consequently, we do not need to estimate the object ordering graph as in [36].

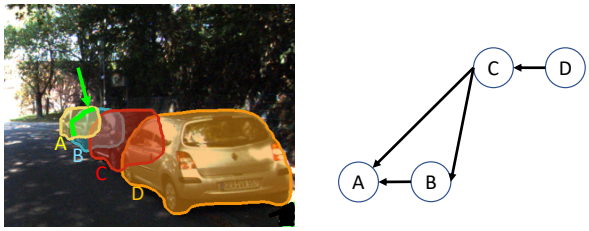

Figure 2. An example of a occluder partially occluded by other objects. A is occluded by B and C, B is occluded by C, and C is occluded by D. For input in the first stage of our training, we use only the occlusion boundary (green), whereas [36] first estimates the ordering graph of A, B, C, and D (on the right), and then takes a union mask of B, C, and D as input.

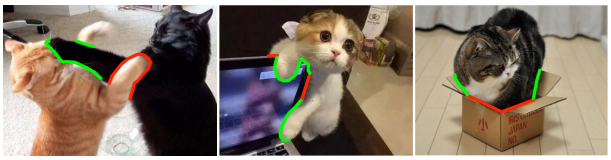

Figure 3. Examples of entangled objects occluding each other, where the occlusion relationships cannot be uniquely represented by an ordering graph, and thus are very challenging for PCNet [36]. As shown in Fig. 5, we successfully address these cases.

Recently, [35] proposes to learn shape priors for each category by using modal object bounding boxes. Then the amodal object bounding box is obtained by aligning modal box with learned shape priors. This approach only works well with low-deformation object categories such as car and motorbike of KINS [31] so that we can robustly learn object shape priors. In contrast, ASBU can handle many types of object category as in COCOA [37].

Uncertainty Estimation in Segmentation has a long track record in the literature. Prior work typically estimates aleatoric uncertainty (data uncertainty) [21] and epistemic uncertainty (weight uncertainty) [16], where the former estimates noise in observations and the latter accounts for a distribution of model parameters. For example, in [21], UNet [33] is extended with a variational auto-encoder for aleatoric uncertainty estimation. In [16], estimation of a distribution of the SegNet parameters [1] replaces the common fixed-point parameter estimation. Shape priors have also been studied in the following related work [18, 19, 25, 34].

3. Our Approach

The section specifies our ASBU. Fig. 4 shows that ASBU uses two distinct sets of inputs for training of UNet to jointly predict the amodal segmentation mask and the associated uncertainty map. These predictions incur an uncertainty-weighted loss function, specified such that our training minimizes both uncertainty and errors in the predicted amodal segmentation.

Our data manipulation of training images, first, randomly samples two objects as occludee and occluder, then, randomly samples a relative position between their modal masks such that the occluder's mask partially occludes the occludee's mask, and finally prepares the following two sets of input data:

1. (set 1 and set 2) Manipulated training image, where the occluder's image is superimposed onto the occludee's image with matting for realistic appearance;
2. (set 1 and set 2) Occlusion boundary mask, estimated as an intersection of the morphologically enlarged masks of the occludee and occluder;
3. (set 1) Occluded mask of the occludee, where pixels of the ground-truth modal mask of the occludee covered by the occluder are zero.
   (set 2) Ground-truth modal mask of the occluder.

Importantly, while both input sets are used in training, ASBU is not aware if the the input segmentation mask comes from the occludee or from the occluder. In this way, ASBU is trained to identify when and how to perform amodal completion of the input segmentation mask. Specifically, for set 1 at the input, ASBU is supposed to learn *to extend* the input segmentation mask in the region with zero pixels, which is delineated by the input occlusion boundary, because this region is likely to represent the manipulated occlusion. On the other hand, for set 2 at the input, ASBU is supposed to learn *not to extend* the input segmentation mask in the zero-valued region.

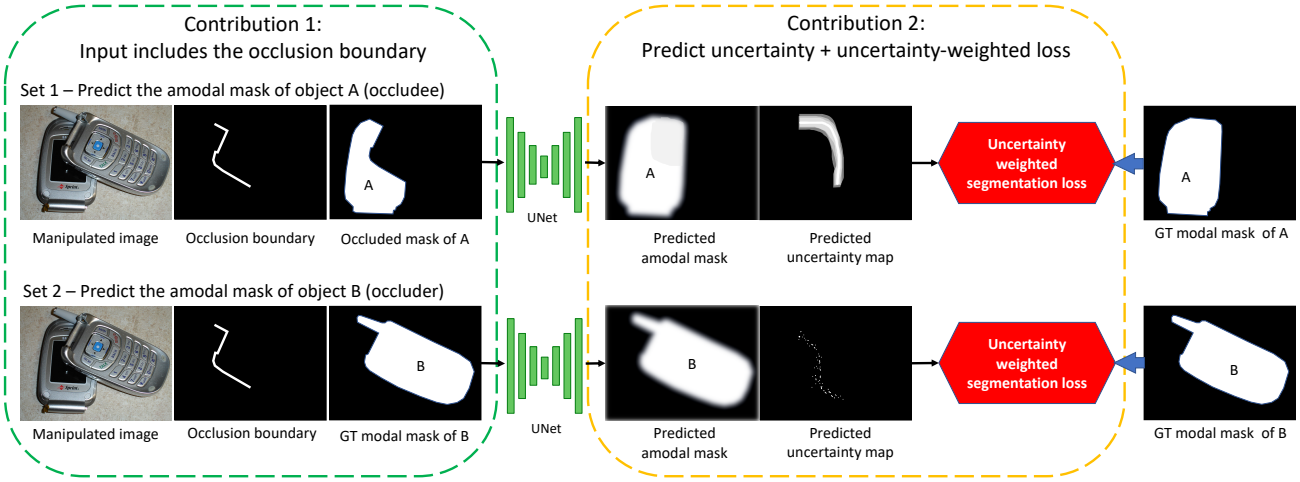

Figure 4. ASBU is trained on two sets of input triplets so as to learn when to spatially extend and when not to extend the input segmentation mask in prediction. Our first contribution is in the input to ASBU– namely, we replace the occluder mask used in [36] with the occlusion boundary mask. Our second contribution is in the prediction of the uncertainty map and using uncertainty to appropriately weight loss. The figure shows that the predicted uncertainty is usually low on regions close to the occlusion. This is used for regularizing learning. Note that we adjusted brightness for visualizing uncertainty (the brighter pixels in the uncertainty map the higher uncertainty), because it is significantly lower for set 2 than for set 1.

4. Uncertainty Weighted Segmentation Loss

Our uncertainty estimation is aimed at implicitly capturing a "shape prior" of training object instances, which is used for regularizing our learning. Thus, our learning has the following two objectives:

1. Minimizing uncertainty of the predicted amodal segmentation, so when uncertainty is estimated as large it captures a truly large variability in plausible shapes;
2. Penalizing a difference between the predicted amodal segmentation and the ground truth in an adaptive manner, such that this loss is appropriately reduced when the shape is estimated to come from a highly variable distribution – i.e., when the estimated uncertainty is high.

To this end, our ASBU extends UNet to output a $H \times W \times 2$ feature map which has two channels, one for amodal segmentation prediction and the another for the uncertainty prediction, where H and W denote the height and width of the input image. The predicted values for amodal segmentation, $\{m_i : i = 1, \ldots, N\}$, $N = H \cdot W$, are output by the sigmoid function, so they range in $m_i \in [0, 1]$. The predicted uncertainty values, $\{u_i : i = 1, \ldots, N\}$, are output by the softplus function, $\text{softplus}(z) = \log(1 + \exp(z))$, so they are positive $u_i \in \mathbb{R}^+$.

For learning to jointly predict $\{m_i\}$ and $\{u_i\}$, we specify a new uncertainty-weighted segmentation loss. The prediction of $\{m_i\}$ can be supervised by the corresponding original modal mask as it was before the data manipulation $\{m_i^*\}$. On the other hand, the prediction of $\{u_i\}$ remains unsupervised, since there is no ground-truth annotation of uncertainty in our training data. We integrate the mentioned supervised and unsupervised training strategies in the following loss:

$$
\begin{aligned}
L &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(m_i^C = 0) L_i + \lambda \mathbb{1}(m_i^C = 1) L_i, \\
L_i &= \frac{1}{2} \left[\left(\frac{m_i^* - m_i}{u_i} \right)^2 + u_i^2 \right],
\end{aligned} \tag{1}
$$

where m^C is the mask of the occluder, λ is a positive constant, $\mathbb{1}(\cdot)$ is the indicator function, and N is the number of pixels in the image. We empirically estimate that $\lambda = 5$ gives the best results, i.e., we put more weight on predictions inside the occluder mask.

Our loss in (1) is inspired by the Mumford-Shah energy [30], whose minimization is a classical framework for image segmentation. The Mumford-Shah energy has two terms – namely, data term and regularization term. By minimizing a sum of these two terms, the predicted segmentation is encouraged to simultaneously minimize energy and complexity (e.g., favor solutions with a high log-likelihood and smooth boundaries).

Similarly, our loss L_i in (1) also has two terms. The first term can be interpreted as the data term for minimizing the energy of a weighted difference between the predicted amodal segmentation m_i and ground truth m_i^*. The weighting is inversely proportional to the estimated uncertainty for pixels i, such that a lower loss is backpropagated in our training for pixels with higher uncertainty. Since u_i is typically higher on object boundaries than over other object parts, the specified weighting effectively accounts for

shape variability. The second term in L_i can be interpreted as the regularization term for penalizing large u_i values. In our experiments, we observe that this regularization favors zero u_i values over object regions that are not close to the boundary. Consequently, the data term in L_i for such regions introduces a large loss for any errors in the amodal segmentation m_i, because $u_i \approx 0$.

Our loss formulation fundamentally differs from other recent approaches aimed at estimating uncertainty for object segmentation. For example, in [17], segmentation and its uncertainty are assumed as governed by a Gaussian distribution with the following loss function:

$$L_i^{\text{Gaussian}} = \frac{1}{2}\left[\frac{(m_i^* - m_i)^2}{u_i^2} + \log u_i^2\right] \tag{2}$$

In contrast, we do not explicitly specify any probability distribution of box locations. Also, unlike our L_i in (1), L_i^{Gaussian} in (2) minimizes uncertainty only when $u > 1$. In [28], uncertainty is modeled in the latent space of their variational auto-encoder, and the amodal mask is predicted by sampling multiple latent codes from a Gaussian distribution. In contrast, we predict and regularize our uncertainty map directly in the spatial domain.

5. Results

Datasets: We evaluate ASBU on COCOA [37] and KINS [31], which are two benchmark amodal instance segmentation datasets with real images derived from the MSCOCO [27] and KITTI [12] datasets, respectively. COCOA consists of 2500, 1323, and 1250 images for training, validation, and test, respectively. There are 2140 object categories that can be divided into two superclasses: stuff (e.g., sky, grass, sea) and things (e.g., dog, cat, human). The content of images is mostly dense with multiple objects occluding one another in cluttered scenes. On the other hand, KINS is a large-scale traffic dataset, which consists of 7474 images for training and 7517 images for testing. There are 7 object categories in KINS including: cyclist, pedestrian, car, tram, truck, van, and miscellaneous vehicles. Scenes in KINS images are less cluttered than in COCOA, and if objects are occluded the occlusion is by mostly one other object. We further randomly divide the KINS training set into training and validation sets with 6000 and 1474 images, respectively. Both datasets provide the ground-truth amodal masks, which we use only for evaluation.

Evaluation Tasks and Metrics: We evaluate ASBU on three tasks: ordering recovery, amodal completion, and amodal instance segmentation.

For ordering recovery, we estimate the following relationships. Let $(\mathbf{m}_j, \mathbf{m}_j^A)$ and $(\mathbf{m}_k, \mathbf{m}_k^A)$ denote two pairs of input modal and output amodal masks of two *adjacent* objects j and k, respectively, where $|\mathbf{m}_j^A - \mathbf{m}_j|$ and $|\mathbf{m}_k^A - \mathbf{m}_k|$ are the extended areas after amodal segmentation of j and k. Then, we specify the ordering of j and k as

$$O(j,k) = \begin{cases} 0, & \text{if } \left|\mathbf{m}_j^A - \mathbf{m}_j\right| = \left|\mathbf{m}_k^A - \mathbf{m}_k\right| = 0 \\ 1, & \text{if } \left|\mathbf{m}_j^A - \mathbf{m}_j\right| < \left|\mathbf{m}_k^A - \mathbf{m}_k\right| \\ -1, & \text{otherwise}, \end{cases} \tag{3}$$

where $O(j,k) = 1$ indicates that j occludes k. If j and k are not adjacent, we set $O(j,k) = 0$. We evaluate our performance on the task of ordering recovery in terms of the average pairwise accuracy, O-Acc, between our predicted ordering relationships $O(j,k)$ and the ground truth relationships $O^*(j,k)$, for all pairs (j,k) of adjacent objects.

For amodal completion, we compute the mean intersection-over-union, mIOU, between the predicted and ground-truth amodal masks, as well as invisible mIoU, invmIoU, for the predicted and ground-truth occluded regions.

For amodal instance segmentation, we report the common metrics suggested by COCO, including average precision AP for thresholds 50%, 75%, 95%, and average recall AR for top 1, 10, 100 predictions, among others.

We evaluate the following baseline and ablations:

- PCNet-m: our strong baseline from [36].
- Boundary→PCNet-m: in the input to PCNet-m we replace the occluder mask with the occlusion boundary; this tests only our contribution 1 (see Fig. 4), as PCNet-m does not estimate uncertainty.
- Uncertainty→PCNet-m: PCNet-m is extended to predict uncertainty and trained with our uncertainty weighted loss, given by (1), while the input uses the occluder mask as in [36]; this tests our contribution 2 (see Fig. 4).
- uBCE→ASBU: the uncertainty weighted loss, given by (1), is replaced with the following uncertainty weighted binary cross-entropy (uBCE) loss for training our ASBU; this tests our proposed data term in (1):

$$L_i^{\text{uBCE}} = \frac{1}{2}\left[-\frac{m_i^* \log m_i + (1-m_i^*)\log(1-m_i)}{u_i^2} + u_i^2\right]. \tag{4}$$

- ASBU: our full approach illustrated in Fig. 4.

5.1. Implementation Details

Our implementation uses the github code of [36] as the base code and modify UNet [32] so it outputs two channels for amodal segmentation and uncertainty map, as described in Sec. 4. We have also tested other networks for segmentation, such as DeepLabv3 [6] and DeepLabv3+ [7]; however, our performance gain using these networks is statistically insignificant. We use the same training setting for fair comparison. For learning, we use SGD with momentum [8], and set the learning rate to $1e^{-4}$. The number of training

Methods	COCOA-val		COCOA-test		KINS-test		
	O-Acc	mIOU	O-Acc	mIoU	O-Acc	mIoU	inv-mIoU
Amodal-VAE [28] (reported)	-	-	-	-	-	94.68	62.85
PCNet-m [36] (reported)	87.10	81.35	-	-	92.50	94.76	-
PCNet-m (reproduced)	85.75	80.73	86.73	86.63	91.73	94.52	59.24
Boundary→PCNet-m	89.01	82.85	89.22	88.67	92.26	94.65	62.77
Uncertainty→PCNet-m	88.60	82.49	88.40	88.15	92.08	94.61	62.00
uBCE→ASBU	89.23	83.18	89.32	88.10	92.15	94.34	63.41
ASBU	**90.33**	**84.22**	**90.77**	**89.87**	**92.65**	**94.83**	**64.41**

Table 1. Evaluation on the tasks of amodal completion and ordering recovery. For comparison with [36], we present the results of PCNet-m (reported) and PCNet-m (reproduced), where the former results are reported in [36], and the latter are obtained by retraining their public code from scratch. '-' indicates that results are not reported.

iterations is 56000 and 32000 for COCOA and KINS, respectively. In each training iteration, we randomly choose Case 1 or Case 2 input data, as described in Sec. 3, to train ASBU with a Bernoulli probability equal to 0.8. The λ in Eq. (1) is set to 5 for the best performance. The threshold to binarize the amodal mask from our network's output is 0.5. The batch size for training UNet is 32 256×256 images.

For amodal instance segmentation on test data, we use Mask-RCNN [13] with ResNet50 [14] as backbone and FPN [26] as the neck. The implementation of Mask-RCNN is provided in mmdetection [4] toolbox. The batch size for training Mask-RCNN is 2 with the default setting provided by mmdetection. All experiments are run on a PC with two 1080 Titan GPUs and 64 GB RAM.

5.2. Ordering Recovery and Amodal Completion

Tab. 1 evaluates ASBU on amodal completion and ordering recovery. For amodal completion, on COCOA-val, both Boundary→PCNet-m and Uncertainty→PCNet-m improve performance in mIoU over PCNet-m by 2.1% and 1.8% over PCNet-m (reproduced), respectively. A similar performance gain is observed on COCOA-test. Our contribution 1 (i.e., using the occlusion boundary mask in the input) has a larger effect on the performance than our contribution 2 (i.e., uncertainty), and each individual contribution leads to performance gains relative to the baseline. The proposed integration of the two contributions in ASBU gives the best amodal completion on both COCOA-val and COCOA-test.

For KINS-test, on amodal completion, ASBU improves performance over PCNet-m (reproduced) in both mIoU and inv-mIoU. The performance gain in mIoU is relatively modest, which can be explained by certain properties of the dataset. KINS has fewer object categories than COCOA (7 vs. 2140), and KINS scenes have fewer occlusions. Therefore, the ordering graph produced by PCNet-m is already highly accurate for amodal completion on KINS.

For the ordering recovery task on COCOA-val and COCOA-test, Tab. 1 shows a similar trend. ASBU significantly outperforms PCNet-m (reproduced).

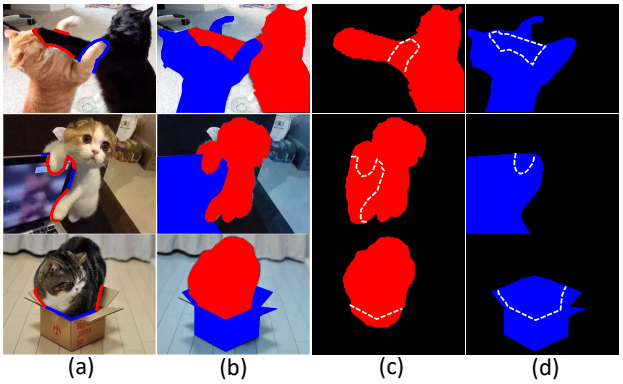

Figure 5. Mutual occlusion examples. ASBU can successfully handle these cases while prior work [36] simply does not work since we cannot define the ordering graph. Each row is an example where red and blue colors represent objects A and B. For each row, from left to right: (a) input RGB image with color boundary segments indicating which object is in front of, (b) modal masks of the two objects, (c) and (d) ASBU predicted amodal masks with white dash line indicating extended regions of objects A and B respectively.

Fig. 5 shows representative examples of two objects that mutually occlude each other. PCNet-m cannot handle such cases since their ordering graph is not expressive enough. On the contrary, ASBU successfully infers the amodal masks of the two mutually occluding objects.

Fig. 6 illustrates results of ASBU on COCOA and KINS on the task of amodal completion. As can be seen, ASBU gives highly accurate predictions. The figure also shows our estimated uncertainty maps which usually take high values on object boundaries. In some cases, ASBU fails to fully complete the amodal masks, due to the high similarity of foreground and background.

5.3. Amodal Instance Segmentation

For this task, we take the trained ASBU to predict amodal masks on COCOA-train and KINS-train, and use these pseudo amodal masks to train Mask-RCNN (pre-

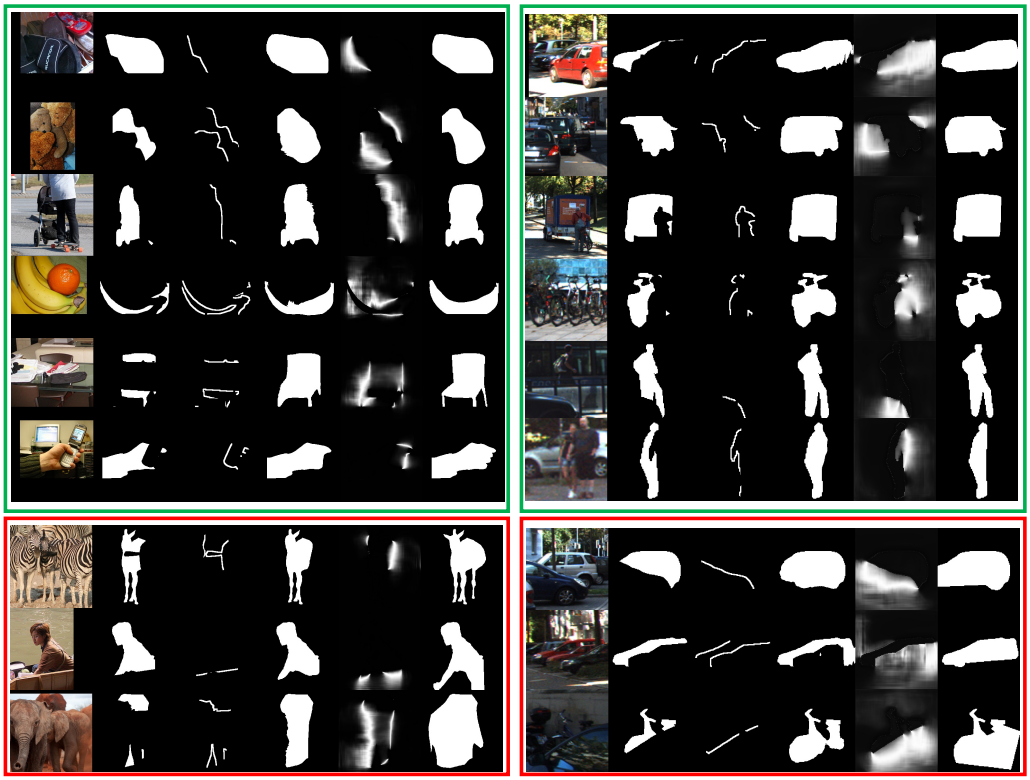

Figure 6. Qualitative results for amodal completion on COCOA-test (left) and KINS-test (right). For each column from left to right: (1) input RGB image, (2) input modal mask, (3) input occlusion boundary, (4) predicted amodal mask, (5) predicted uncertainty map, (6) GT amodal mask. Successful cases are in the green bounding boxes, and failure cases are in the red bounding boxes.

Datasets	**Trained on**	**AP**	**AP$_{50}$**	**AP$_{75}$**	**AP$_S$**	**AP$_M$**	**AP$_L$**	**AR$_1$**	**AR$_{10}$**	**AR$_{100}$**	**AR$_S$**	**AR$_M$**	**AR$_L$**
COCOA-val	GT amodal	22.2	44.8	20.0	13.8	20.6	24.3	6.0	27.4	39.3	33.4	39.4	40.0
	PCNet-m amodal	21.0	43.4	18.5	13.7	19.5	22.9	5.9	26.6	37.9	33.8	38.6	38.0
	ASBU amodal	22.2	44.5	20.0	12.5	19.8	24.6	6.1	27.4	38.9	33.1	39.1	39.5
COCOA-test	GT amodal	23.9	48.4	21.5	14.1	23.0	25.8	6.4	28.7	40.9	31.7	41.6	41.5
	PCNet-m amodal	22.6	46.8	19.7	13.7	22.0	24.2	6.3	27.7	39.2	32.3	40.0	39.6
	ASBU amodal	23.8	47.9	21.2	13.8	22.4	25.6	6.4	28.6	40.5	32.9	40.9	41.1
KINS-test	GT amodal	30.8	53.9	31.6	15.1	40.4	56.7	18.9	38.3	40.4	24.1	51.6	65.6
	PCNet-m amodal	29.1	51.8	29.6	14.1	38.1	55.7	18.3	37.1	38.9	23.0	49.4	65.2
	ASBU amodal	29.3	52.1	29.7	14.2	38.2	56.0	18.4	37.0	38.8	23.1	49.3	64.9

Table 2. Amodal instance segmentation results of Mask-RCNN in full COCO metrics on COCOA-val, COCOA-test and KINS-test. Mask-RCNN is trained on either GT amodal masks, or PCNet-m generated amodal masks or ASBU generated amodal masks.

trained on COCO) in 12 epochs (1x configuration) to predict amodal instance segmentation for COCOA-val, COCOA-test, and KINS-test. We use the evaluation code from the COCO dataset Github. We repeat the same process for trained PCNet-m. Because the number of classes in COCOA is too large (2140 classes) and our focus is on the quality of amodal segmentation, we group them into one foreground class to train and evaluate. For KINS, we keep the number of classes as specified in this dataset.

For reporting an upper-bound performance, we train Mask-RCNN on the ground-truth amodal masks of COCO-train and KINS-train to predict amodal instance segmentation of COCOA-val, COCOA-test, and KINS-test.

Tab. 2 evaluates amodal instance segmentation using Mask-RCNN trained on: ground-truth amodal segmentations (GT amodal), and pseudo-ground truth produced by PCNet-m (PCNet-m amodal) and ASBU (ASBU amodal). From the table, on COCOA-val, a difference in AP between GT amodal and ASBU amodal is zero. On COCOA-test, when using the pseudo-ground truth from ASBU, we increase AP relative to that of PCNet-m amodal. Tab. 2 suggests that on COCOA ASBU pseudo amodal masks have

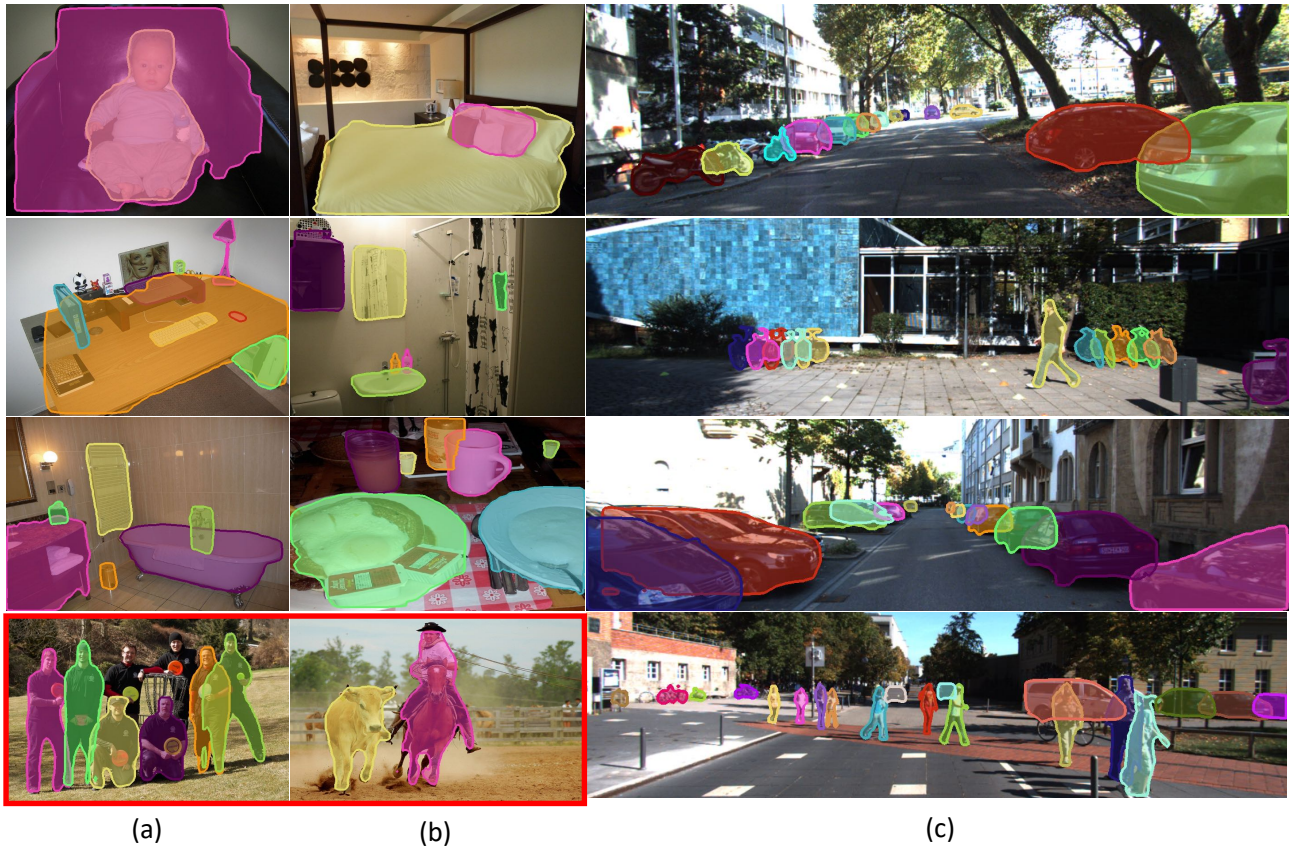

Figure 7. Qualitative results for amodal instance segmentation on COCOA-test are shown in (a) and (b), and on KINS-test in (c). All of them are successful cases except the failure cases are marked red. The first failure case is about incomplete person detections and the second failure case is about merging masks of a person and horse.

similar quality as the actual ground truth. On KINS-test, ASBU amodal gives slightly better results than PCNet-m amodal (with a 0.2 margin) while we are behind from GT amodal by 1.5 in AP. This can be explained in terms of simpler scenes in KINS relative to those in COCOA.

Fig. 7 shows representative results of Mask-RCNN trained with ASBU's pseudo amodal masks. For the COCOA-test dataset, we usually obtain good results, with some exceptions due to the problems of incomplete instance modal segmentation. Also, on the KINS-test dataset, we obtain very good amodal instance segmentations.

6. Conclusion

We have specified a new amodal segmenter with boundary uncertainty estimation (ASBU) for weakly supervised amodal instance segmentation. To address the lack of ground-truth amodal masks, we have trained ASBU on manipulated images to produce pseudo-ground truth amodal masks, and then learned a common instance segmenter, Mask-RCNN, on our pseudo-ground truth. We have made two contributions. First, we have replaced the occluder mask used in prior work [36] for input with the occlusion boundary, and consequently removed the need for one step in [36] – that of estimating the object ordering graph. Second, we have enabled ASBU to estimate uncertainty of the predicted amodal segmentation and proposed a new loss function that uses the estimated uncertainty to regularize learning of ASBU. Our evaluation on the tasks of amodal completion, ordering recovery, and amodal instance segmentation, on the COCOA dataset, demonstrates that ASBU outperforms the state of the art. Specifically, in comparison with a strong baseline PCNet-m [36], our performance improves by 3.5% and 4.5% in mean intersection-over-union (mIoU) for amodal completion and average pairwise accuracy (O-Acc) for ordering recovery, respectively. In amodal instance segmentation, Mask-RCNN trained on our pseudo amodal masks has nearly the same performance as Mask-RCNN trained on the ground-truth amodal masks with a performance gap of 0.1 in average precision (AP) on the COCOA-test dataset.

Acknowledgement. This work was supported in part by DARPA MCS Award N66001-19-2-4035.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 3

[2] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: real-time instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9157–9166, 2019. 2

[3] Siddhartha Chandra, Nicolas Usunier, and Iasonas Kokkinos. Dense and low-rank gaussian crfs using deep embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5103–5112, 2017. 2

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[5] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2018. 2

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 5

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *The European Conference on Computer Vision (ECCV)*, September 2018. 5

[8] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 5

[9] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6144–6153, 2018. 2

[10] Alireza Fathi, Zbigniew Wojna, Vivek Rathod, Peng Wang, Hyun Oh Song, Sergio Guadarrama, and Kevin P Murphy. Semantic instance segmentation via deep metric learning. *arXiv preprint arXiv:1703.10277*, 2017. 2

[11] Patrick Follmann, Rebecca Kö Nig, Philipp Hä Rtinger, Michael Klostermann, and Tobias Bö Ttger. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1328–1336. IEEE, 2019. 2

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 5

[13] 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, 6

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 6

[15] Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, and Alexander G Schwing. Sail-vos: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3105–3115, 2019. 2

[16] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015. 3

[17] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017. 5

[18] Yuka Kihara, Matvey Soloviev, and Tsuhan Chen. In the shadows, shape priors shine: Using occlusion to improve multi-region segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 392–401, 2016. 2, 3

[19] Benjamin B Kimia, Ilana Frankel, and Ana-Maria Popescu. Euler spiral for shape completion. *IJCV*, 54(1-3):159–182, 2003. 3

[20] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother. Instancecut: from edges to instances with multicut. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5008–5017, 2017. 2

[21] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Ali Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. In *Advances in Neural Information Processing Systems*, pages 6965–6975, 2018. 3

[22] 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. 2

[23] Ke Li and Jitendra Malik. Amodal instance segmentation. In *ECCV*, pages 677–693. Springer, 2016. 2

[24] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2359–2367, 2017. 2

[25] Hongwei Lin, Zihao Wang, Panpan Feng, Xingjiang Lu, and Jinhui Yu. A computational model of topological and geometric recovery for visual curve completion. *Computational Visual Media*, 2(4):329–342, 2016. 3

[26] 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. 6

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva

Ramanan, C. Lawrence Zitnick, and Piotr Dollr. Microsoft coco: Common objects in context. In *European conference on computer vision (ECCV)*, June 2016. 5

[28] Huan Ling, David Acuna, Karsten Kreis, Seung Kim, and Sanja Fidler. Variational amodal object completion for interactive scene editing. In *NeurIPS*, 2020. 3, 5, 6

[29] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[30] David Bryant Mumford and Jayant Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 1989. 4

[31] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2019. 2, 3, 5

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 2, 5

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, 2015. 3

[34] Nathan Silberman, Lior Shapira, Ran Gal, and Pushmeet Kohli. A contour completion model for augmenting surface reconstructions. In *ECCV*, 2014. 3

[35] Yihong Sun, Adam Kortylewski, and Alan Yuille. Weakly-supervised amodal instance segmentation with compositional priors. *arXiv preprint arXiv:2010.13175*, 2020. 3

[36] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene deocclusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3784–3792, 2020. 1, 2, 3, 4, 5, 6, 8

[37] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1464–1472, 2017. 2, 3, 5