2019 IEEE Winter Conference on Applications of Computer Vision

Learning to See the Invisible: End-to-End Trainable Amodal Instance Segmentation

Patrick Follmann Rebecca König Philipp Härtinger Michael Klostermann Tobias Böttger
MVTec Software GmbH
follmann@mvtec.com

Abstract

Semantic amodal segmentation is a recently proposed extension to instance-aware segmentation that includes the prediction of the invisible region of each object instance. We present the first all-in-one end-to-end trainable model for semantic amodal segmentation that predicts the amodal instance masks as well as their visible and invisible part in a single forward pass.

In a detailed analysis, we provide experiments to show which architecture choices are beneficial for an all-in-one amodal segmentation model. On the COCO amodal *dataset, our model outperforms the current baseline for amodal segmentation by a large margin. To further evaluate our model, we provide two new datasets with ground truth for semantic amodal segmentation,* D2S amodal *and* COCOA cls*. For both datasets, our model provides a strong baseline performance. Using special data augmentation techniques, we show that amodal segmentation on* D2S amodal *is possible with reasonable performance, even without providing amodal training data.*

1. Introduction

Humans, with their strong visual system, have no difficulties reasoning about foreground and background objects in a two-dimensional image. At the same time, humans have the ability of *amodal perception*, *i.e.* to reason about the invisible, occluded parts of objects [10, 11]. Robots that should navigate in their environment and pick or place objects need to know if the objects are occluded or hidden by one or several other instances. This problem leads to the task of semantic amodal segmentation, *i.e.*, the combination of segmenting each instance within an image by predicting its amodal mask and determining which parts of the segmented instances are occluded and what the corresponding occluder is. A typical example is shown in Figure 1.

The amodal mask is defined as the union of the visible mask (which we will also refer to as modal mask) and the invisible occlusion mask of the object (*c.f.* Figure 1). Predicting amodal and visible masks simultaneously provides a deeper understanding of the scene. For example, it allows to calculate regions of occlusion and lets the robot know which objects have to be removed or in which direction to move in order to get free access to the object of interest.

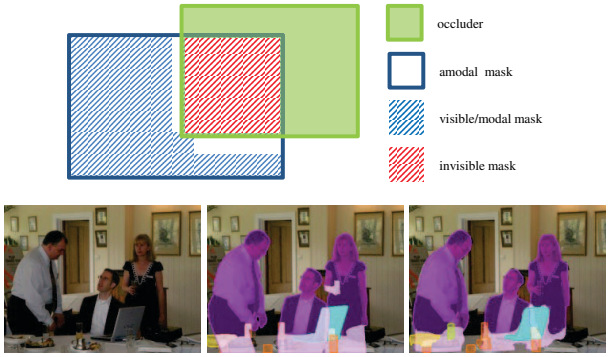

Figure 1. **Learning the Invisible.** *Top:* explanation of different mask types; *Bottom, from left to right* input image, ground-truth amodal instance annotations, predictions of our model: amodal instance masks and occlusion masks. The mask color encodes the object class and occlusion masks are highlighted in light color.

Predicting the invisible part of an object is difficult: If the object is occluded by an object from another category, the model has no visual cues how to extend the visible mask into the occluded object part. There are generally no edges or other visual features that indicate the contour of the occluded object. In contrast, if the object is occluded by another instance of the same category, it is very hard for the model to judge where to stop expanding the mask into the occlusion part as the category-specific features are present all around.

We propose a model that can predict the visible, invisible, and amodal masks for each instance simultaneously without much additional computational effort. In summary, our paper contains the following contributions:

- To the best of our knowledge, we are the first to propose an all-in-one, end-to-end trainable multi-task

978-1-7281-1975-5/19/$31.00 ©2019 IEEE
DOI 10.1109/WACV.2019.00146

model for semantic segmentation that simultaneously predicts amodal masks, visible masks, and occlusion masks for each object instance in a single forward pass.

- We provide the new semantic amodal segmentation dataset *D2S amodal*, which is based on *D2S* [3], with guaranteed annotation-completeness and high-quality annotations. In comparison to the class-agnostic *COCO amodal* dataset [19], *D2S amodal* contains 60 different object categories and allows to predict amodal and occlusion masks class-specifically.

- By merging the categories of the modal *COCO* dataset with the instances of *COCO amodal* we obtain the new amodal dataset *COCO amodal cls* with class labels.

- Our architecture ORCNN outperforms the current baseline on *COCO amodal* [19] and sets a strong baseline on *D2S amodal*. We provide extensive evaluations in order to compare different architectural choices.

- The training set of *D2S* allows to apply extensive data augmentation. This allows to train a semantic amodal method without any amodally annotated data. The model achieves competitive results on *D2S amodal*.

Note that throughout the paper, we will call annotations containing only visible masks and models predicting visible masks *modal*, in contrast to *amodal* annotations and methods. We will also use the terms occlusion masks and invisible masks as synonyms.

2. Related Work

The topic of amodal perception has already been addressed in various fields of computer vision research.

Semantic Segmentation and 3D Scene Reconstruction. Two tasks, where amodal completion has already been used for some years, are semantic segmentation and 3D scene reconstruction. The task of semantic segmentation is to predict a category label for each pixel in an image. Semantic segmentation does not take different object instances into account, but returns a single region for each of the possible classes. Classes are often related to background or *stuff*, such as *sky, water, ground, wall, etc*. In [5], Guo and Hoiem describe a method to infer the entire region of occluded background surfaces. Their algorithm detects the occluding objects and fills their regions with the underlying or surrounding surface.

In 3D reconstruction, parts of the scene can often not be reconstructed because of occlusions. Gupta *et al*. [6] combine depth information, superpixels, and hierarchical segmentations for amodal completion of semantic surfaces. Also Silberman *et al*. [17] address the problem of surface completion in the setting of a 2.5D sensor. They use a conditional random field in order to complete contours. The completed contours are subsequently used for surface completion.

In contrast to the above mentioned semantic segmentation methods, our work does not deal with the amodal completion of background regions or 3D object surfaces, but focuses on object instances in 2D images.

Object Detection. In the context of object detection, Kar *et al*. [9] use a CNN to predict amodal bounding boxes of objects. By additionally estimating the depth of the bounding boxes and the focal length of the camera, object dimensions can be derived. However, neither the object mask nor the occluded part of the object is predicted.

Instance Segmentation. More recent methods extend the *object detection* task to the more challenging *instance segmentation* task to predict the category and visible segmentation mask of each object instance in an image [14, 7, 12]. Yang *et al*. [18] propose a probabilistic model that uses the output of object detectors to predict instance shapes and their depth ordering. However, no occlusion regions are predicted. In [1], Chen *et al*. propose a graph-cut algorithm with occlusion handling in order to improve the quality of visible masks. However, they neither predict occlusion nor amodal masks.

Amodal Instance Segmentation. Research on amodal instance segmentation or semantic amodal segmentation has just started to emerge. Li and Malik [13] were the first to provide a method for amodal instance segmentation. They extend their instance segmentation approach [12] by iteratively enlarging the modal bounding box of an object into the directions of high heatmap values and recomputing the heatmap. Due to the lack of amodal instance segmentation ground truth, they use modally annotated data and data augmentation in order to train and evaluate their model.

In [19], Zhu *et al*. provide a new and pioneering dataset *COCO amodal* for amodal instance segmentation based on images from the original *COCO* [15] dataset. The authors did not restrict the annotations to the usual *COCO* classes and annotators could assign arbitrary names to the objects. Therefore, all objects in the dataset belong to a single class *object* and the variety of objects in this class is very large. Additionally, the authors provide annotations of background regions, which are sometimes extending to the full image domain, labeled as *stuff*. In order to provide a baseline, Zhu *et al*. use AmodalMask, which is the SharpMask [16] model trained on the amodal ground truth. The model suggests object candidates with a relatively high recall. However, the predictions of the model are class-agnostic. They also trained a ResNet-50 [8] to predict the

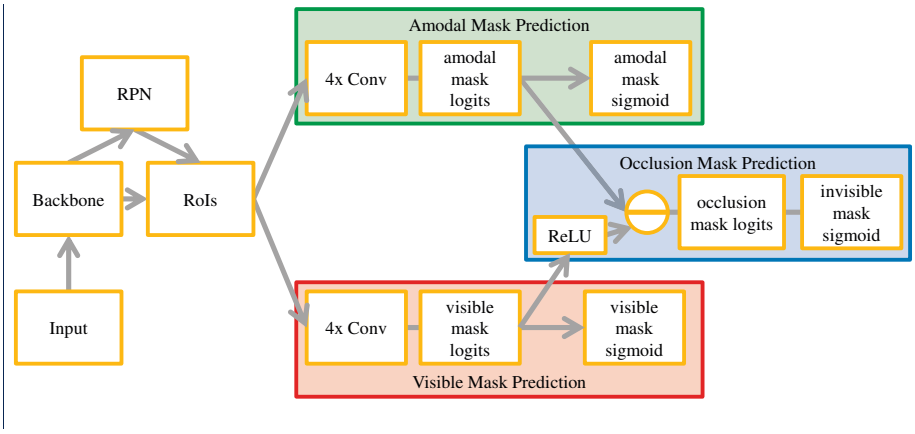

Figure 2. **ORCNN architecture.**

foreground object given two input object masks and the corresponding image-patches.

In contrast to [13] and [19], our model is class-specific, end-to-end trainable, lightweight, and can predict amodal, visible, and invisible instance masks in a single forward-pass.

3. End-to-End Architecture for Prediction of Amodal, Visible, and Invisible Masks

3.1. Architecture

We name our method *Occlusion R-CNN* (*ORCNN*), as the architecture is based on Mask R-CNN [7] (MRCNN). In the ORCNN architecture we extend MRCNN with additional heads for the prediction of amodal masks (amodal mask head) and the occlusion mask (occlusion mask head). An overview of the ORCNN architecture is shown in Figure 2.

The visible mask head and the amodal mask head share the same architecture and use four 3×3 convolutions and ReLU layers to generate meaningful features for mask prediction. Their inputs are the extracted features from the RoIAlign [7] layer. Note that during training and inference, the amodal and visible mask prediction heads of ORCNN share the same box proposals generated by the region proposal network (RPN). The target ground truth masks of the RPN are the bounding boxes of the amodal instances. Therefore, the visible mask prediction head has to predict the visible mask of an instance from the amodal bounding box. This is a major difference to a modal model that is trained using the bounding boxes of the modal, *i.e.* visible masks of the instances.

A key component is that we link the modal and amodal mask heads with the occlusion mask head. The occlusion mask head essentially subtracts the visible from the amodal mask logits in order to obtain the occlusion mask logits. It is crucial to apply a ReLU-operation on the visible mask logits before subtraction to avoid occlusion mask prediction for pixels where neither the amodal nor the modal mask are predicted.

Mounting both the modal, as well as the amodal head on the same RoI-feature extraction module leads to several advantages: First, this makes the additional amodal and occlusion mask prediction light-weight as only five additional convolution modules and two sigmoid layers are necessary. Second, using the same RoIs for the amodal and visible mask prediction guarantees that both predicted masks correspond to the same object prediction. In comparison, if one uses an ensemble of two separate models for amodal and visible mask predictions, it is not straightforward to fuse the results of these models. And third, by sharing the same architecture for the amodal and visible mask prediction head, we can initialize both heads with the same weights that have been pre-trained on a large modal instance-segmentation dataset, such as *COCO*.

3.2. Training

In order to obtain meaningful predictions for the visible, invisible, and occlusion masks, we have to formulate the corresponding losses for each of the tasks. As the tasks are similar and only the ground truth differs, we use a similar sigmoid-cross-entropy loss for all three mask types: First applying a per-pixel sigmoid and thereafter an average binary cross-entropy loss like in [7]. In combination with the losses for the class (*cls*) and bounding box (*box*), we obtain the total loss L:

$$L = L_{cls} + L_{box} + L_{AM} + L_{VM} + L_{IVM}, \qquad (1)$$

where *AM, VM, IVM* are abbreviations for amodal, visible, and invisible mask, respectively. In theory, one of the three losses L_{AM}, L_{VM} and L_{IVM} is redundant, as for ground truth we have $IVM = AM - VM$. Nevertheless, adding an additional loss for occlusion masks leads to amodal mask logits and visible mask logits that are on the same scale.

Otherwise, consider the case that for a pixel we have high probability for the amodal and the visible mask. E.g., let the logit activations be 14 and 10, respectively. This leads to a probability of $1/(1+e^{-(14-10)}) = 0.982$ for the occlusion mask to be present at this pixel, although occlusion should not be predicted.

In order to test the influence of adding the visible and invisible mask losses, we experiment with four different model variants: standard ORCNN is using a loss for each of the three mask types (amodal, visible and invisible). For ORCNN (w/o L_{IV}) and ORCNN (w/o L_V) the loss for invisible or visible mask prediction is switched off, respectively. ORCNN (independent) is a model including L_{IV} and L_V but where the gradients with respect to L_{IV} and L_V are not propagated to the amodal mask head nor to the RoI feature extraction part.

As an alternative to include the invisible mask prediction by the model, we will also show results where the invisible masks are computed as the difference between the amodal and visible mask outputs. However, for ORCNN, the direct prediction of invisible masks comes at negligible cost.

3.3. Evaluation

To judge which model is the best for the task of amodal instance segmentation, we propose extending the mean average precision (AP) and mean average recall (AR) evaluation measures commonly used for instance segmentation, e.g. on *Pascal VOC* [2] and *COCO* [15] benchmarks. For brevity, in the following we will describe the extension of the measure only for the case of AP. The extension for AR is straight-forward. As in the *COCO* benchmark, we compute the final AP by taking the mean of the per-category APs. As is common practice, the per-category APs are averaged over ten equally spaced intersection over union (IoU) thresholds from 0.5 to 0.95 to highlight more precise results.

In order to evaluate the individual tasks, we calculate the AP values independently for amodal and visible masks to obtain AP_A and AP_V, respectively. We can then include both of the masks into the definition of a true positive instance to obtain a combined AP measure AP_{AV} (amodal-visible AP). For example, in order to obtain a true positive result in the AP_{AV} setting, for a given IoU threshold t, we need the correct predicted class and additionally, $\mathrm{IoU}(AM^G, AM^P) > t$ and $\mathrm{IoU}(VM^G, VM^P) > t$ both need to be satisfied. Here, AM^G and AM^P denote the amodal mask ground truth and the amodal mask prediction AM^P, respectively.

The invisible masks are included only indirectly into the overall measure AP_{AV} due to the following issues: First, for non-occluded objects, the invisible mask is not present and it is not straight-forward to define recall on something that is not present[1]. Second, for most objects in *COCO amodal* or *D2S amodal*, the invisible mask areas are rather small compared to the amodal or visible masks. Hence, small differences in invisible mask predictions have a large influence on the IoU.

To measure the quality of predicted occlusions, we do separate evaluations, where we ignore all non-occluded ground truth objects. For these evaluations, we also calculate the average precision of invisible masks $AP_{IV}^{0.5}$ at IoU-threshold 0.5. We use a low IoU threshold as the invisible masks are often of very small size and to take the difficulty of the task into account. For the calculation of AP_{IV}, we only use results where the amodal segmentation has an IoU higher than 0.5 with a ground truth amodal mask of an occluded object. I.e., if the predicted amodal mask is a false positive, we ignore the corresponding invisible mask.

For models that do not predict any visible or invisible masks, we calculate the measures AP_{AV} and AP_V by using amodal masks also as predictions for the visible masks.

When invisible masks are calculated as the difference of amodal and visible masks, we denote the corresponding measure AP_{IV} *diff*.

4. Experiments

In the following, we compare our models to previous results on *COCO amodal* (*COCOA*) and set new benchmarks for the new semantic amodal datasets *COCO amodal cls* and *D2S amodal*. All models were trained using the Detectron [4] framework. More information on the settings is in the supplementary material.

4.1. *COCOA*

COCOA [19] is the first dataset with ground truth for semantic amodal segmentation. The dataset consists of 2500 training, 1323 validation, and 1250 test images. In each image, most objects and background *stuff* areas are annotated with amodal masks. Occluded objects are additionally annotated with visible and invisible masks. All objects belong to a single category *object* and have an additional *stuff* label.

Amodal Mask Prediction. As a baseline result for amodal semantic segmentation we train MRCNN with a ResNet-50 or ResNet-101 backbone on the amodal annotations. We call these models *ARCNN-50/ARCNN*. Table 1 compares ARCNN to the baseline AmodalMask of [19] using their evaluation tool. ARCNN outperforms AmodalMask by a large margin in terms of average precision. AmodalMask achieves a high recall, since it always predicts 1000 regions for each image. Nevertheless, ARCNN

[1] Generally, recall is increasing with more proposals. In case of an empty invisible mask this is not the case, which prevents the usual way of computing the AP measure.

	all		things		stuff	
	AP_A	AR_A	AP_A	AR_A	AP_A	AR_A
AmodalMask [19]	5.7	43.4	5.9	45.8	0.8	**36.7**
ARCNN-50	**29.9**	**45.8**	**33.2**	**50.4**	**5.8**	33.0

Table 1. **Baseline results for *COCOA*.** Amodal mean average precision and average recall values for AmodalMask [19] compared to ARCNN.

achieves even higher recall while predicting only 30 results per image on average. An exception is the category *stuff*, since *stuff* regions are often very large areas in the background of the image and ARCNN predicts no object proposals for these areas.

In the following, our focus is on *things*, as masks for *stuff* are hard to define and therefore, the variance of annotations between different annotators is high. Thus, in *COCOA no stuff*, we exclude *stuff* annotations during training and evaluation. We found that for ARCNN and ORCNN AR is generally in line with AP since recall is already captured within the AP measure. Therefore, for the following evaluations we will just show AP values. In order to highlight the performance on occluded objects, we also evaluate the architectures ignoring all non-occluded instances.

Occlusion Prediction. As a baseline for a model like ORCNN that can predict amodal, visible, and invisible masks at the same time, we use ARCNN and standard MRCNN. Since both models do only predict amodal or visible masks, respectively, in the evaluation, we use the amodal masks also as visible mask predictions and vice-versa.

To combine the benefits of MRCNN and ARCNN, we also use an ensemble-approach applying both models and merging the results (mergedAMRCNN). Therefor, we match the modal results to the amodal results greedily: We merge the modal result with highest IoU to the amodal result (if $\mathrm{IoU} > 0.5$) if the predicted classes match. The score of a match is set to the mean of amodal and visible mask scores. If an amodal result is not matched, the amodal mask is used as visible mask and the invisible mask is set to an empty region. The best results are obtained if the unmatched modal results are ignored (thrown away).

The results on *COCOA* are summarized in the top two sections of Table 2.The multi-task model ORCNN improves the quality of visible masks compared to ARCNN and sometimes even compared to MRCNN, while at the same time predicting occlusion masks. The best result for the combined measure AP_{AV} is obtained using the variant ORCNN (independent). This variant combines the benefits of the ensemble mergedAMRCNN into a single model with only a slight performance decrease for some of the measures. Standard ORCNN is the best choice for the prediction of invisible masks. Qualitative results for *COCOA* are shown in the supplementary material.

4.2. *COCOA cls*

For amodal completion the model has to get some intuition about the common shape of objects. We evaluate whether the prediction of amodal masks is a class-specific task. Therefore, we generated a new dataset *COCOA cls* by merging the usual *COCO* 2014 annotations with the *COCOA* dataset. *COCOA* contains many objects of categories (e.g. sandals, sneakers, or stuff categories) that are not part of *COCO*. Although, each object in *COCOA* has a name tag, the annotators were free to choose a name. Furthermore, not all objects present in the ground truth of *COCO* have been annotated in *COCOA*. Therefore, to assign a class-label to the objects in *COCOA*, we calculate the IoUs of the visible masks with the masks given for the corresponding image-id in *COCO*. Only annotations, for which the IoU between visible mask and any *COCO* annotation exceeds a threshold of 0.75 (and not labeled as stuff or crowd), were kept for *COCOA cls*. Overall, *COCOA cls* has 3,501 images with 10,592 objects compared to the 3,823 images and 34,916 objects of *COCOA*. Note that using this merging scheme, some *COCO* classes, e.g. *hairdryer*, are not present in the training set of *COCOA cls*. Furthermore, for many images not all *COCO* annotations can be matched to a corresponding *COCOA* label.

As shown in Table 2, models perform a lot better on *COCOA cls* compared to *COCOA*. On *COCOA cls* the ensemble mergedAMRCNN is the best choice, slightly increasing the performance compared to the multi-task-model ORCNN (independent). The ensemble is in favor, as it averages the scores of both predictions: Only if both models are confident, the final score is high. Also false positive predictions of MRCNN are filtered out in mergedAMRCNN if there is no corresponding false positive of ARCNN. This reduces the total number of false positives.

In order to see if class-specific mask prediction improves the results, ARCNN and ORCNN were also trained using class-agnostic mask proposals. On *COCOA cls*, using class-specific mask proposals helps for almost all measures both in the case of ARCNN and ORCNN. For occluded objects, class-agnostic ORCNN gets the best average precisions in terms of invisible masks.

Generally, the interpretation of AP values for *COCOA* and *COCOA cls* results is difficult because for both datasets the annotations are not complete. For example, as shown in Figure 3, there are some cases where ORCNN makes correct predictions but the ground truth does not contain the corresponding annotations.

4.3. *D2S amodal*

D2S [3] is a recent dataset of supermarket products that should capture the setting and needs of industrial appli-

dataset & model	all			occluded				
	AP_{AV}	AP_A	AP_V	AP_{AV}	AP_A	AP_V	$AP_{IV}^{0.5}$	$AP_{IV}^{0.5}$ diff
COCOA								
AmodalMask [19]	3.7	5.7	4.8	1.2	3.2	2.6	-	-
MRCNN [7]	17.0	18.6	21.7	8.6	10.6	15.3	-	-
ARCNN	24.1	**31.3**	26.1	11.8	**21.6**	16.2	-	-
mergedAMRCNN	22.5	28.5	27.0	12.5	19.5	18.5	-	0.3
ORCNN	21.2	25.7	26.7	11.5	17.0	18.4	**1.3**	**1.7**
ORCNN (independent)	**25.0**	31.1	**28.8**	**13.7**	**21.6**	**19.9**	1.0	0.5
COCOA no stuff								
MRCNN [7]	22.0	23.9	27.9	12.4	15.0	21.6	-	-
ARCNN	27.8	**35.6**	29.4	14.7	**25.4**	18.5	-	-
mergedAMRCNN	28.1	33.6	**33.6**	**16.8**	23.5	**24.4**	-	0.4
ORCNN	25.1	30.1	30.0	14.3	20.8	21.4	**3.0**	**1.9**
ORCNN (independent)	**29.0**	35.1	32.8	16.6	25.0	23.0	1.0	0.5
COCOA cls								
MRCNN [7]	39.0	39.8	44.9	25.4	26.5	34.9	-	-
ARCNN (agn)	37.1	40.4	38.6	23.9	28.7	26.8	-	-
ARCNN	38.8	41.7	40.5	24.9	29.2	28.0	-	-
mergedAMRCNN	**40.1**	**42.5**	**45.7**	**27.7**	**30.0**	**34.9**	-	1.0
ORCNN (agn)	34.3	36.2	39.3	23.1	25.0	29.5	1.8	1.4
ORCNN	35.1	37.6	39.4	23.7	25.8	29.9	**2.0**	1.4
ORCNN (independent)	38.0	40.7	41.0	26.0	28.9	30.5	0.2	1.0

Table 2. ***COCOA* results.** Note that only ORCNN is predicting visible and invisible masks additionally to the amodal masks. For all other models, the predicted amodal mask was used for the evaluation of AV and V measures. Models marked with (agn) are using class-agnostic mask prediction heads.

Figure 3. ***COCOA cls* results.** *top:* ground truth annotations for images of the validation set; *bottom:* exemplary qualitative results of our ORCNN model trained on the `train` split. Predicted invisible masks are indicated by a white overlap. Note that ORCNN sometimes predicts instances or invisible masks correctly that are not part of the ground truth. These count as false positives in the evaluation or at least lead to reduced AP_{AV} and AP_{V} values.

cations. In particular, the low complexity of the training set with respect to its size and image attributes (homogeneous background, no clutter) makes it necessary to use data augmentation. Since only minor occlusions are present in the training set, the generation of reasonable artificial images is straight-forward. We annotated all images of *D2S* amodally to obtain *D2S amodal*. The annotations contain the category, amodal mask, and additional visible and invisible masks for occluded objects. For images where the amodal masks are reaching out of the image boundary, a zero-padding is used such that all amodal masks are fully contained in the image.

Splits. As the *D2S amodal* training split only contains minor occlusions, it is not suitable to train semantic amodal models that predict occlusion masks also for moderately to heavily occluded objects. Hence, we use data augmentation similar to the data augmentation in [3]. *D2S amodal augmented* consists of 1562 augmented images, where only objects from *D2S train* that do not reach out of the image boundary have been used for augmentation. The *D2S*

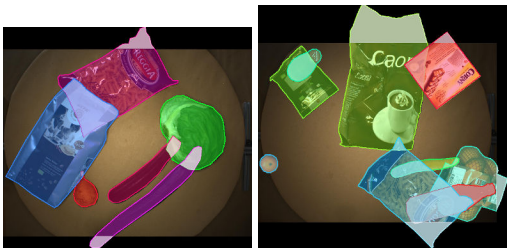

Figure 4. **D2S amodal ground truth.** Exemplary ground truth annotations for different splits of the *D2S amodal* dataset. *(left)* val, *(right)* aug.

amodal train set is then the combination of *D2S amodal training_rot0* with *D2S amodal augmented* splits resulting in a total of 2000 images.

To evaluate if an amodal model can be trained only from modal annotated data, we augmented another 2000 images in the same way, but using the modal annotations from *D2S training* to obtain *D2S amodal modal augmented*. For all splits of *D2S amodal*, the statistics are given in Table 3. As is common for *D2S*, the number of images containing occlusion are a lot higher for the validation set compared to the train_rot0 set. Using the augmented set, one can create an overall train split that contains many objects and even a higher frequency and average per-object occlusion rate than in the validation and test splits. Exemplary ground truth annotations are shown in Figure 4.

split name	aug	train_rot0	train	val	modal aug
num imgs	1562	438	2000	3600	2000
num imgs w/ occl	1507	57	1564	2520	1930
img OR [%]	96	13	78	70	96
num objs	12376	690	13066	15654	15851
num objs occl	8798	66	8864	7473	11302
obj OR [%]	71	9	68	47	71
avg OR / reg (all) [%]	23	0	22	8	23
avg OR / reg (occl) [%]	33	4	33	18	33

Table 3. ***D2S amodal* splits.** Image and occlusion statistics for the splits used with *D2S amodal* from top to bottom: number of images, number of images with at least one occluded object, rate of images that contain any occlusion, total number of objects, number of occluded objects, rate of objects that are occluded, average occlusion rate per object region for all objects, average occlusion rate per object only for occluded objects.

Results. The qualitative results of ORCNN in Figure 5 are very promising. The model predicts occlusions correctly in many cases, especially for objects lying completely on top of another one or objects reaching out of the image boundary. This shows that the model is able to learn the common shape of object classes.

The quantitative analysis given in Table 4 shows that the prediction of invisible masks is very difficult. On the one hand, if the ground truth invisible masks are small, then small differences of the invisible mask proposal already lead to a low IoU value. On the other hand, if the ground truth invisible mask is large, it is very difficult for the model to generate the correct shape of the invisible mask prediciton, again leading to a low IoU value. In both cases if the IoU value is below 0.5 this leads to a false positive prediction and possibly a false negative as the corresponding ground truth is not matched.

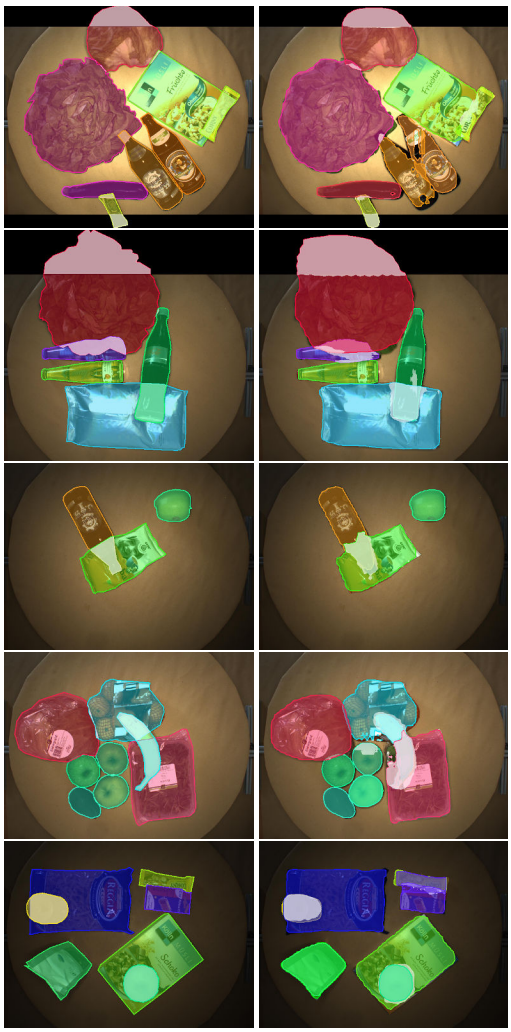

Figure 5. ***D2S amodal* results.** *(left)* ground truth annotations for images of the validation set; *(right)* exemplary qualitative results of our ORCNN model trained on the `amodal_train` split. Predicted invisible masks are indicated by a white overlap. In most cases the results are promising, especially for objects lying on top of each other or reaching beyond the original image boundary. More results are in the supplementary material.

On *D2S amodal*, ORCNN (independent) achieves comparable results to mergedAMRCNN and even outperforms the ensemble with respect to AP_{AV} on occluded objects.

dataset & model	all			occluded				
	AP_{AV}	AP_A	AP_V	AP_{AV}	AP_A	AP_V	$AP_{IV}^{0.5}$	$AP_{IV}^{0.5}$ diff
ARCNN (agn)	63.4	72.2	64.8	48.0	62.2	50.8	-	-
ARCNN	63.8	72.6	65.3	48.7	**63.0**	51.6	-	-
MRCNN (agn)	64.7	65.2	**77.7**	48.7	49.3	**71.3**	-	-
mergedAMRCNN (agn)	**69.2**	**72.7**	75.5	58.2	62.8	68.2	-	13.2
ORCNN (w/o L_{IV})	34.0	68.6	34.7	26.8	58.9	28.0	0.1	0.0
ORCNN (w/o L_V)	11.0	66.1	11.1	7.5	56.7	7.6	5.8	0.0
ORCNN (agn)	62.3	65.2	71.0	52.5	55.2	65.3	**14.7**	**13.8**
ORCNN	58.9	62.1	66.2	48.1	51.2	58.9	8.7	8.9
ORCNN (independent, agn)	**69.2**	72.3	74.3	**59.1**	62.5	67.6	0.8	9.2
ORCNN (modal aug)	56.3	59.9	62.6	47.9	51.0	57.6	7.8	-

Table 4. ***D2S amodal* results.** Models marked with (agn) are using class-agnostic mask prediction heads. ORCNN (modal aug) is a ORCNN only trained on *D2s amodal modal augmented*. All values are calculated for the *D2S amodal validation* set.

Figure 6. **D2S failure cases.** Exemplary qualitative results where our ORCNN model fails. Predicted invisible masks are indicated by a white overlap. *(top)* occlusion is falsely predicted possibly due to reflections or lighting changes; *(bottom-left)* the amodal and invisible mask of this instance is extended into the neighboring object of the same class (only this instance is visualized); *(bottom-right)* the occlusion caused by the plastic bottle is not detected.

For invisible mask prediction, standard ORCNN is the best choice. Interestingly, ORCNN gets a higher $AP_{IV}^{0.5}$ when the invisible mask is predicted directly instead of using the difference between amodal and visible mask.

Some failure cases of ORCNN on *D2S amodal* are shown in the appendix. False positive occlusion predictions are often caused by reflections or lighting changes. When objects are lying next to each other and touching, the amodal and invisible masks are sometimes extended into the neighboring instance. For other cases, occlusions are not predicted at all.

In summary, ORCNN (independent) yields the best compromise for the prediction of visible and invisible masks at the same time (highest AP_{AV} and AP_V). This comes at the cost of a slightly lower AP_V value compared to a model that only predicts visible masks, like MRCNN. Note that compared to ORCNN, mergedAMRCNN has a much higher memory consumption (2×7600 MB vs. 8300 MB), as well as significantly longer inference runtime (2×170 ms vs 180 ms, nvidia GTX1080Ti GPU). Furthermore, there is an additional overhead for the offline merging strategy (≈ 10 ms), which depends on the number of output results.

In contrast to *COCOA*, for *D2S amodal* ORCNN (agn) outperforms the ORCNN model that predicts a class specific mask.

Table 5 also shows the result of ORCNN when training only on artificially augmented data obtained from *D2S train*, *D2S amodal modal augmented*. The model performs only slightly worse than ORCNN trained on *D2S amodal train*. Thus, it is possible to train a competitive model without the need of amodally annotated data.

5. Conclusion

We proposed an end-to-end trainable, instance-aware model for semantic amodal segmentation. Our model, ORCNN, simultaneously predicts amodal, visible and invisible masks, and the category of each instance in a single forward pass. By merging annotations of *COCO* with *COCOA*, we obtain a category-specific semantic amodal dataset based on *COCO* images *COCOA cls*. We provide semantic amodal ground truth for *D2S* splits resulting in *D2S amodal*. ORCNN (independent) outperforms previous work on *COCOA* and sets strong benchmarks on *COCOA cls* as well as *D2S amodal*. The lightweight, all-in-one model is able to achieve comparable performance to an ensemble approach. Furthermore, we show that it is possible to train a competitive amodal model only using modal annotations and data augmentation. Our experiments and results show that it is possible to predict the invisible masks of occluded objects even in areas without any visual cue. Thus, our model can indeed *learn to see the invisible*.

The results show that the prediction of amodal and in particular invisible masks is a difficult task that needs further research to reduce the number of false positive predictions of occlusions.

References

[1] Y.-T. Chen, X. Liu, and M.-H. Yang. Multi-instance object segmentation with occlusion handling. In *Conference on Computer Vision and Pattern Recognition*, pages 3470–3478, 2015. 2

[2] M. Everingham, S. M. A. Eslami, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 4

[3] P. Follmann, T. Böttger, P. Härtinger, R. König, and M. Ulrich. MVTec D2S: Densely Segmented Supermarket Dataset. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision – ECCV 2018*, pages 581–597, Cham, 2018. Springer International Publishing. 2, 5, 6

[4] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He. Detectron. https://github.com/facebookresearch/detectron, 2018. 4

[5] R. Guo and D. Hoiem. Beyond the line of sight: labeling the underlying surfaces. In *European Conference on Computer Vision*, pages 761–774. Springer, 2012. 2

[6] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Conference on Computer Vision and Pattern Recognition*, pages 564–571. IEEE, 2013. 2

[7] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN. In *International Conference on Computer Vision*, pages 1059–1067, 2017. 2, 3, 6

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2

[9] A. Kar, S. Tulsiani, J. Carreira, and J. Malik. Amodal completion and size constancy in natural scenes. In *International Conference on Computer Vision*, pages 127–135, 2015. 2

[10] P. J. Kellman and C. M. Massey. Perceptual learning, cognition, and expertise. In *Psychology of learning and motivation*, volume 58, pages 117–165. Elsevier, 2013. 1

[11] S. Lehar. Gestalt isomorphism and the quantification of spatial perception. *Gestalt theory*, 21:122–139, 1999. 1

[12] K. Li, B. Hariharan, and J. Malik. Iterative instance segmentation. In *Conference on Computer Vision and Pattern Recognition*, pages 3659–3667, 2016. 2

[13] K. Li and J. Malik. Amodal instance segmentation. In *European Conference on Computer Vision Submission*, pages 677–693. Springer, 2016. 2, 3

[14] Y. Li, H. Qi, J. Da, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In *Conference on Computer Vision and Pattern Recognition*, pages 2359–2367, 2017. 2

[15] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision Submission*, pages 740–755, 2014. 2, 4

[16] P. O. Pinheiro, R. Collobert, and P. Dollar. Learning to segment object candidates. In *Advances in Neural Information Processing Systems 28*, pages 1990–1998. 2015. 2

[17] N. Silberman, L. Shapira, R. Gal, and P. Kohli. A contour completion model for augmenting surface reconstructions. In *European Conference on Computer Vision*, pages 488–503. Springer, 2014. 2

[18] Y. Yang, S. Hallman, D. Ramanan, and C. C. Fowlkes. Layered object models for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1731–1743, 2012. 2

[19] Y. Zhu, Y. Tian, D. Metaxas, and P. Dollar. Semantic amodal segmentation. In *Conference on Computer Vision and Pattern Recognition*, pages 1464–1472, 2017. 2, 3, 4, 5, 6