Amodal Instance Segmentation

Ke Li(✉) and Jitendra Malik

Department of Electrical Engineering and Computer Sciences,
University of California, Berkeley, USA
{ke.li,malik}@eecs.berkeley.edu

Abstract. We consider the problem of amodal instance segmentation, the objective of which is to predict the region encompassing both visible and occluded parts of each object. Thus far, the lack of publicly available amodal segmentation annotations has stymied the development of amodal segmentation methods. In this paper, we sidestep this issue by relying solely on standard modal instance segmentation annotations to train our model. The result is a new method for amodal instance segmentation, which represents the first such method to the best of our knowledge. We demonstrate the proposed method's effectiveness both qualitatively and quantitatively.

Keywords: Instance segmentation · Amodal completions · Occlusion reasoning

1 Introduction

Consider the horse shown in the left panel of Fig. 1. The task of instance segmentation requires marking the visible region of the horse, as shown in the middle panel, and has been tackled by several existing algorithms [6,7,18,19,27]. In this paper, we consider a different task, which requires marking both the visible and the occluded regions of the horse, as shown in the right panel. In keeping with terminology used in the psychology literature on visual perception [20], we refer to the former task as *modal instance segmentation* and the latter task as *amodal instance segmentation*.

A natural question to ask is if the task of amodal instance segmentation is well-posed: given only the visible portions of an object, there are many possible configurations of the hidden portions of the object, all of which appear to be plausible hypotheses to a human. This is particularly true for articulated objects under heavy occlusion. For example, if the lower body of a person is blocked from view, there is no single correct hypothesis for the configuration of the person's legs – the person could be sitting or standing, and so hypotheses consistent with either pose would be equally valid. Despite this ambiguity, humans are capable

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-46475-6_42) contains supplementary material, which is available to authorized users.

B. Leibe et al. (Eds.): ECCV 2016, Part II, LNCS 9906, pp. 677–693, 2016.
DOI: 10.1007/978-3-319-46475-6_42

Fig. 1. Target outputs for modal and amodal segmentation

of performing amodal completion and tend to predict the occluded regions with high degrees of consistency [40].

An amodal segmentation system would open the way to sophisticated occlusion reasoning. For instance, given an amodal segmentation mask, we can infer the presence, extent, boundary and region of occlusions by comparing it to the modal segmentation mask. We can also deduce the relative depth ordering by comparing the modal and amodal masks of the occluded and occluding objects. The information derived from the amodal segmentation mask can be further used downstream for a variety of interesting applications. For instance, we can estimate the physical dimensions of an object in the real world using its amodal bounding box, as demonstrated by Kar et al. [21].

The fact that all these occlusion reasoning problems can be reduced to amodal instance segmentation implies that amodal instance segmentation is more challenging than all these problems combined. An amodal segmentation system must not only be capable of determining if an object is occluded, but also where it is occluded. It must be able to hypothesize the shape of the occluded portion even though it has never seen the whole object before. It must be sensitive enough to detect the diminished signal from the small part of the occluded object that remains visible, but must be robust enough to avoid being misled by strong signals from occluding objects.

Furthermore, additional complicating factors arise if one were to attempt training a model for the task, the chief among them being the lack of supervised training data. While efforts are underway to collect amodal segmentation annotations [40], no amodal segmentation data is publicly available at the time of writing. In the meantime, we need to devise a clever way of using existing data to train a model for this new task.

In this paper, we present a new method for amodal instance segmentation, which to the best of our knowledge is the first such method. We train our method purely from existing modal instance segmentation data, thereby sidestepping issues arising from the lack of supervised training data. We make a key observation: while it is not possible to compute the amodal mask of an object from the modal mask by undoing occlusion, it is easy to do the reverse. Instead of undoing existing occlusion, we add synthetic occlusion and retain the original mask, which essentially becomes the true amodal mask for the composite image. We

train a convolutional neural net to recover the original mask from the generated composite image. We do not assume knowledge of the amodal bounding box at test time; instead, we infer it from the amodal segmentation heatmap using a new strategy, which we dub Iterative Bounding Box Expansion. We demonstrate that despite being trained on synthetic data, the resulting model is quite effective at predicting amodal masks on images with real occlusions.

2 Related Work

Efforts toward understanding the semantic meaning associated with free-form regions in images started with work on figure-ground segmentation, the objective of which is to identify the foreground pixels in typically object-centric images. Early methods [4,23,25,26,38] investigated ways of combining top-down and bottom-up segmentation approaches and incorporating class-specific or object-specific templates of the foreground object's appearance. Later on, the focus shifted to the more general problem of semantic segmentation, which aims to identify the pixels that belong to each object category in more complex images. A diverse range of approaches for this problem have been developed; earlier approaches extend CRF-based formulations [3,22,24], consider combining object detections with region proposals [15], scoring groupings of over-segmented regions [35], aggregating information from multiple foreground-background hypotheses [5] and synthesizing scores from different overlapping regions to obtain a pixel-wise classification [1]. Later approaches [10,29,31,39] use feedforward or recurrent neural net models to extract features or to predict the final label of each pixel directly from images. In effort at achieving understanding of a scene at a finer level of granularity, recent work has focused on the task of instance segmentation, the goal of which is to identify the pixels that belong to each individual object instance. The predominant framework for this task is to find the bounding box of each instance using an object detector and predict the figure-ground segmentation mask inside each box. Earlier approaches rely on DPM [11] detections and predict segmentation masks using a simple appearance model [8,30,37], combine DPM detections and semantic segmentation predictions [9,12] or adopt a transductive approach [34]. More recent methods leverage the power of convolutional neural nets. SDS [18] and Dai et al. [6] use a neural net to compute features on region proposals and classifies them using an SVM, while the Hypercolumn net [19], Iterative Instance Segmentation [27] and Multi-task Network Casades [7] predict the segmentation mask directly from the image patch. We view the task of amodal instance segmentation as the natural next step in this direction.

There has been relatively little work exploring amodal completion. Kar et al. [21] tackled the problem of predicting the amodal bounding box of an object. Gupta et al. [16] explored completing the occluded portions of planar surfaces given depth information. To the best of our knowledge, there has been no algorithmic work on general-purpose amodal segmentation. However, there has been work on collecting amodal segmentation annotations. Zhu et al. [40]

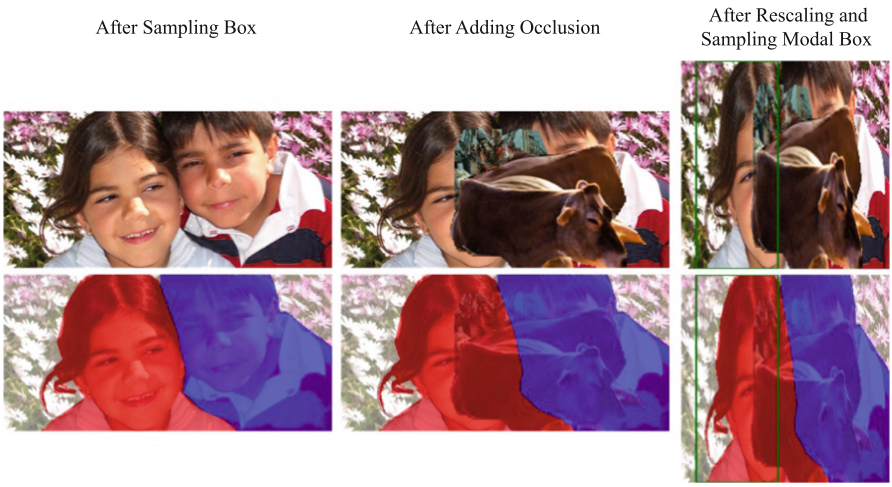


Fig. 2. The image patch and the target segmentation mask after each step of the sampling procedure. Red regions in the segmentation mask are assigned the positive label, white regions in the segmentation mask are assigned the negative label and blue regions in the segmentation mask are assigned the unknown label. The green box denotes the jittered modal bounding box. (Colour figure online)

collected amodal segmentation annotations on BSDS images, but has yet to make them publicly available. As far as we know, the proposed method represents the first method for amodal segmentation.

3 Generating Training Data

3.1 Overview

We generate amodal training data solely from standard modal instance segmentation annotations. In our case, we use the Semantic Boundaries (SBD) annotations [17] on the PASCAL VOC 2012 *train* set as the data source. We generate three types of data: image patches, modal bounding boxes and target segmentation masks. Image patches and modal bounding boxes are used as input to the model and target segmentation masks are used as supervisory signal on the output of the model.

The key observation we leverage is that the phenomenon of occlusion can be easily simulated by overlaying objects on top of other objects. More concretely, we first generate randomly cropped image patches that overlap with at least one foreground object instance, which we will refer to as the main object. We then extract random object instances from other images and overlay them on top of the randomly cropped patches with their modal segmentation masks serving as the alpha matte. Each overlaid object is positioned and scaled randomly in a way that ensures a moderate degree of overlap with the main object. Essentially,

this procedure generates composite patches where the main object is partially occluded by other objects.

Next, for each composite patch, we find the smallest bounding box that encloses the portions of the main object that remain visible. This is essentially ground truth modal bounding box of the main object in the composite patch. To simulate noisy modal localization at test time, we jitter the bounding box randomly.

Finally, we generate the target segmentation masks corresponding to the composite patches produced above. For each patch, we take the corresponding part of its original modal segmentation mask and label the pixels belonging to the object as positive, pixels belonging to the background as negative and pixels belonging to other objects as unknown. This manner of label assignment captures what we know about the amodal mask given the modal mask – we know the visible portion of the object must be a part of the whole object and that the object cannot be occluded by the background. However, the object *may* be occluded by other objects in the image; consequently, the pixels belonging to other objects in the modal mask are labelled as unknown. Because the original modal mask is not affected by overlaid objects, this mask includes portions of the main object that were originally visible but are now occluded in the composite patch. Hence, the target mask is consistent with the true amodal mask. The data generation process is illustrated in Fig. 2 and examples of generated patches and masks are shown in Fig. 3.

3.2 Implementation Details

Data is generated on-the-fly during training. To generate a training example, we sample an image uniformly and then sample an object instance from the image, which will be referred to as the main object. We then randomly sample a bounding box that overlaps with the main object's bounding box along each dimension by at least 70 %. The size of the sampled box is randomly chosen and the length of each dimension is between 70 % and 200 % of the length of the corresponding dimension of the object's bounding box. Next, we choose the number of objects to overlay onto the patch inside the bounding box randomly by picking an integer from 0 to 2. To select an object to overlay, we sample an image and then sample a random object instance from the image. The object is placed at a random location that overlaps with the main object and scaled randomly so that the length of its shortest dimension is 75 % of the length of the corresponding dimension of the patch on average. After each of the above operations, we check if the proportion of the main object that remains visible falls below 30 %. If it does, we undo the most recent operation and try again. Otherwise, we proceed to find the bounding box the encloses the portion of the main object that remains visible and randomly samples a box that overlaps with the bounding box by at least 75 % along each dimension and can differ in size from the bounding box by at most 10 % in each dimension.

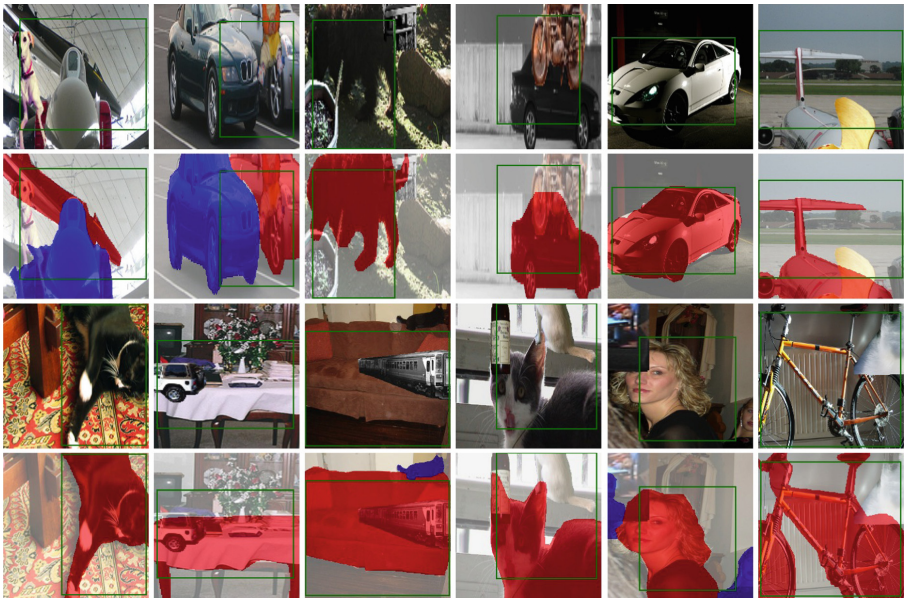


Fig. 3. Random samples from the generated training data

4 Predicting Amodal Mask and Bounding Box

4.1 Testing

We take the modal bounding box, that is, the bounding box of the visible part of the object, and the category of the object as given, which can be obtained from an object detector, like R-CNN [14], fast R-CNN [13] or faster R-CNN [32]. We then compute the modal segmentation heatmap using Iterative Instance Segmentation (IIS) [27], which is the state-of-the-art method for modal instance segmentation.

The proposed algorithm proceeds to predict the amodal segmentation mask and bounding box in an iterative fashion using a new strategy that will be referred to as Iterative Bounding Box Expansion (Fig. 4). Initially, the amodal bounding box is set to be the same as the modal bounding box. In each iteration, given the amodal bounding box, we feed the patch inside the amodal bounding box to a convolutional neural net, which predicts the amodal segmentation mask inside an expanded amodal bounding box that also includes areas immediately outside the original amodal bounding box. We compute the average heat intensity in the areas lie above, below, to the left and to the right of the original bounding box. If the average heat intensity associated with a particular direction is above a threshold, which is set to 0.1 in our experiments, we expand the bounding box in that direction and take this new bounding box to be the amodal bounding box used in the next iteration. This procedure is repeated until the average heat intensities in all directions are below the threshold. To obtain the

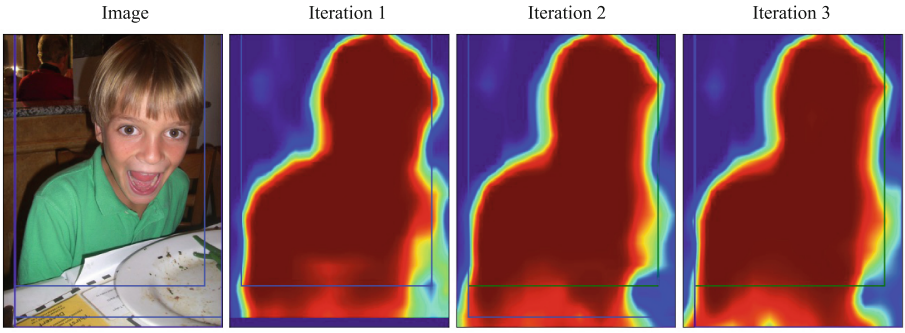


Fig. 4. Amodal segmentation heatmap after each iteration of Iterative Bounding Box Expansion. The green box denotes the modal bounding box and the blue box denotes the amodal bounding box. In each iteration, we use the average heat intensity *outside* the amodal bounding box to decide whether to expand the amodal bounding box in the next iteration. (Colour figure online)

final amodal segmentation mask, we colour in all pixels whose intensities in the corresponding heatmap exceed 0.7. Similarly, we obtain the modal segmentation mask by thresholding the modal heatmap at 0.8.

4.2 Training

The convolutional neural net we use takes in an image patch, a modal segmentation heatmap and a category and outputs the amodal segmentation heatmap. The net has the same architecture as that used by IIS, which is based on the hypercolumn architecture introduced by Hariharan et al. [19]. This architecture is designed to take advantage of both low-level image features at finer scales and high-level image features at coarser scales. It does so by making the final heatmap prediction dependent on the summation of upsampled feature maps from multiple intermediate layers, which is known as the hypercolumn representation. The version of the architecture we use is based on the VGG 16-layer net [33], which is referred to as "O-Net" in [19]. The IIS architecture is a variant of this architecture that also takes in an initial heatmap hypothesis via an additional category-dependent channel as input, which can be set to constant if no initial heatmap hypothesis is available. If an initial heatmap hypothesis is provided, the model refines the heatmap hypothesis to produce an improved heatmap prediction. Using this architecture, IIS is able to iteratively refine its own heatmap prediction by feeding in its heatmap prediction from the preceding iteration as input.

Each training example consists of an image patch, a modal bounding box and a target amodal segmentation mask. To prepare input to the net, we take the part of image patch that lies inside the modal bounding box and scale it anisotropically to 224×224, feed it to IIS as input. We take the modal segmentation heatmap produced by one iteration of IIS, align it to the coordinates

of the original image patch and upsample it to 224×224 using bilinear interpolation. Because the model should predict the mask corresponding to an area larger than the image patch it sees, we remove 10 % of the image patch from each of the four sides and rescale it to 224×224. If less than 10 % of the pixels in this new patch belong to the visible portion of the object, we reject the current sample and generate a new training example. We centre the data by subtracting the mean pixel from the image patch and transforming the modal segmentation heatmap element-wise to lie between -127 and 128. Finally, we feed the image patch, the modal segmentation heatmap and the target amodal mask to the model for training.

The model is trained end-to-end using stochastic gradient descent with momentum on mini-batches of 32 patches starting from weights of the model used by IIS. The loss function we use is the sum of pixel-wise negative log likelihoods over all pixels with known ground truth labels. We apply instance-specific weights that are inversely proportional to the factor by which each patch is upsampled. We train the model with a constant learning rate of 10^{-5}, weight decay of 10^{-3} and momentum of 0.9 for $50,000$ iterations.

5 Experiments

Because there is no existing dataset with amodal instance segmentation annotations, there is no ground truth against which the predictions can be evaluated, making it difficult to perform a quantitative evaluation. We first present qualitative results and then perform an indirect quantitative evaluation against coarse-level annotations on the full PASCAL VOC 2012 *val* set. To perform a direct quantitative evaluation, we annotated 100 randomly chosen occluded objects from the same dataset with amodal segmentation masks and evaluate the proposed method on this subset.

5.1 Qualitative Results

We use the proposed method to generate amodal segmentation mask predictions for objects in the PASCAL VOC 2012 *val* set. Because the focus of this paper is on the segmentation system, we take the modal bounding box and the category of the object of interest as given and obtain them from the ground truth. We show the amodal heatmap and mask predictions produced by the proposed method in Figs. 5, 6 and 7, along with the modal heatmap and mask predictions generated by IIS for comparison. As shown, the proposed method is generally quite effective. In particular, even though the synthetic occlusions generated for training purposes do not appear entirely realistic, the proposed method is able to devise plausible hypotheses for hidden portions of objects caused by real occlusions.

We classify occlusions into two types: cases where the occluding object is mostly contained inside the occluded object and cases where a significant portion of the occluding object lies outside the occluded object. We will refer to the

former as interior occlusions and the latter as exterior occlusions. For interior occlusions, the goal is to predict the mask between visible portions of the object, whereas for exterior occlusions, the goal is to predict the mask beyond the visible portion. There is typically a single correct way to handle interior occlusions: if it results from a true occlusion, the corresponding hole should be filled in; otherwise, it should not be. Therefore, for these cases, the task is less ambiguous and relatively easy. On the other hand, there are generally multiple equally valid ways to handle exterior occlusions: there are many possible ways to extend the visible portion that all lead to plausible amodal masks. So, for these cases, the task is more ambiguous and challenging. The method need to decide how much to extend the modal mask by in every direction. To do so, it must rely on the knowledge it learns about the general shape of objects of the particular category to produce a plausible amodal mask hypothesis.

We first examine examples of occluded objects for which the proposed method produces correct amodal segmentation masks, which are shown in Fig. 5. As shown, the proposed method is able to successfully predict the amodal mask on images with interior or exterior occlusions. Surprisingly, on some images where the modal prediction is poor, like the image with a dog on a folding chair, the proposed method is able to produce a fairly good amodal mask. Finally, the proposed method is able to produce remarkably good amodal masks on some challenging images with exterior occlusions, such as the image depicting a dog on a kayak and the images in the bottom two rows, suggesting that the proposed method is able to learn something about the general shape of objects.

Next, we take a look at examples of occluded objects for which the predicted amodal segmentation masks are incorrect, which are shown in Fig. 6. The mistakes may be caused by the rarity of unusual poses in the training set, large variation in the plausible configurations of the occluded portions of the object, similarity in appearance of adjacent objects or erroneous modal predictions.

While the preceding examples show that the model is capable of performing amodal completion, we need to make sure that the model does not hallucinate when presented with unoccluded objects. We explore some examples of unoccluded objects and the corresponding mask predictions in Fig. 7. Since the objects are unoccluded, amodal masks should be the same as modal masks. As shown, the amodal predictions are similar to or more accurate than the modal predictions. This may be explained by the robustness acquired by an amodal segmentation model: by learning to be robust to occlusion, the model also learns to be robust to variations in low-level patterns in the image that may confuse a modal segmentation model.

We include additional examples of heatmap and mask predictions in the supplementary material.

5.2 Indirect Evaluation

We use the proposed method combined with a modal instance segmentation method like IIS to predict the presence or absence of occlusion. We do so by

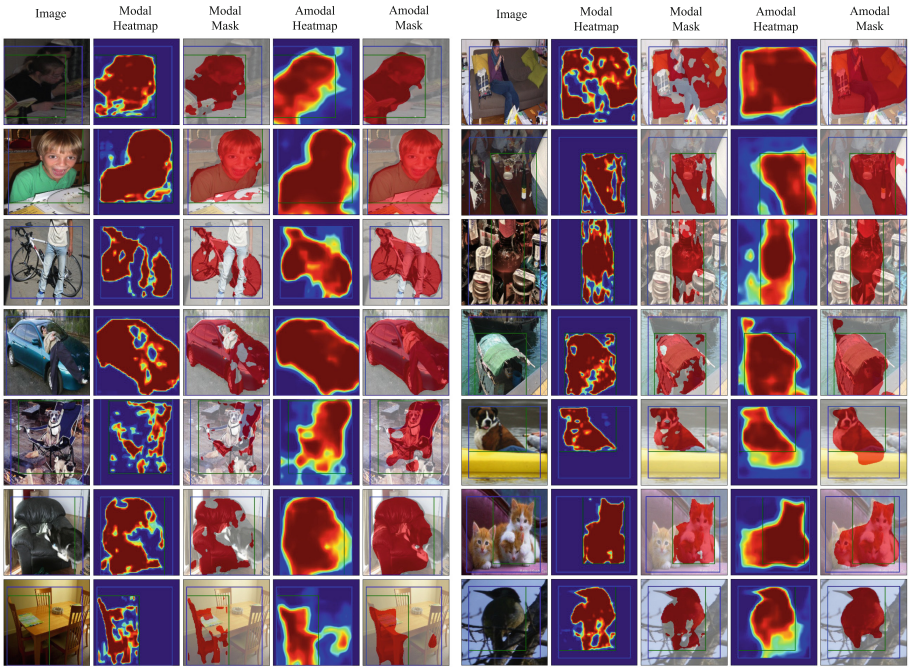


Fig. 5. Examples of amodal segmentation mask predictions for occluded objects which we judge to be correct. The first column shows the part of the image containing the amodal and modal bounding box, with the green box denoting the original modal bounding box that is given and the blue box denoting the expanded amodal bounding box found by Iterative Bounding Box Expansion. If a side of the amodal bounding box is adjacent to the border of the image patch, the patch abuts the corresponding border of the whole image. The next four columns show the modal segmentation heatmap and mask produced by IIS and the amodal segmentation heatmap and mask produced by the proposed method (Color figure online)

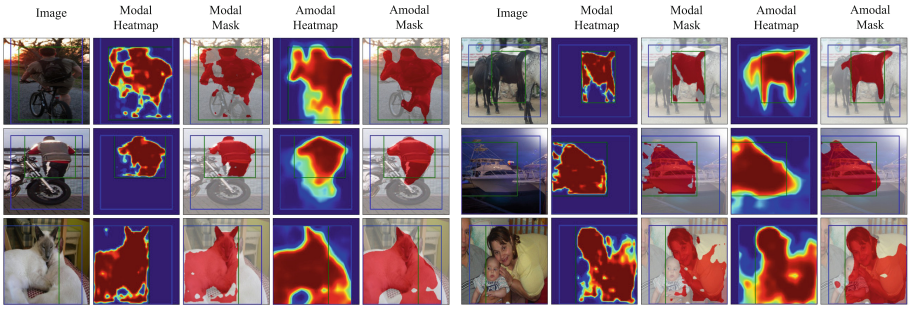


Fig. 6. Examples of amodal segmentation mask predictions for occluded objects which we judge to be incorrect. The visualizations follow the same format as Fig. 5

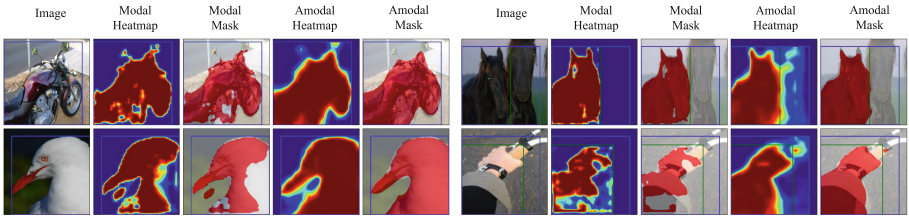


Fig. 7. Examples of amodal segmentation mask predictions for unoccluded objects. The visualizations follow the same format as Fig. 5

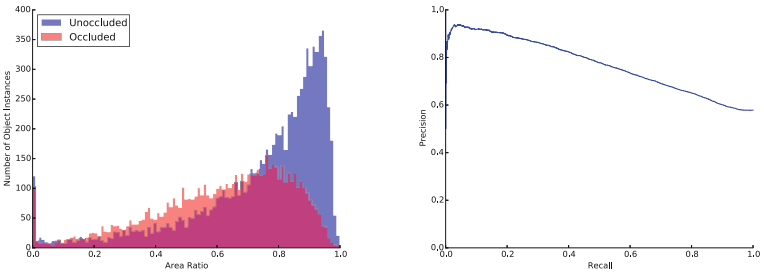
predicting the modal and amodal segmentation masks for each object in the PASCAL VOC 2012 *val* set and computing the following ratio:

$$\frac{\text{area}\,(\textit{modal mask} \cap \textit{amodal mask})}{\text{area}\,(\textit{amodal mask})}$$

Intuitively, this ratio measures the degree by which an object is occluded. For an unoccluded object, because the amodal mask should be the same as the modal mask, this ratio should be close to 1. On the other hand, for a heavily occluded object, only a small proportion of the pixels inside the amodal mask should also be included in the modal mask; as a result, this ratio should be significantly less than 1. We will henceforth refer to this ratio as the *area ratio*.

We compare our predictions to the occlusion presence annotations in the PASCAL VOC 2012 *val* set, which are available for all object instances and specify whether they are occluded. First, we compute the modal and amodal mask predictions using IIS and the proposed method for all objects that are annotated as being occluded and plot the distribution of area ratios. We then do the same for objects that are annotated as being unoccluded. These two distributions are shown in Fig. 8a. As shown, the distribution for unoccluded objects is heavily skewed towards high area ratios, whereas the distribution for occluded objects peaks at an area ratio of around 0.75. This indicates the predicted amodal masks for occluded objects typically have more pixels outside modal masks than those for unoccluded objects, which confirms that the proposed method generally performs amodal completion only for occluded objects and not for unoccluded objects. The distribution for occluded objects is also flatter than that for unoccluded objects because the amount by which an object is occluded by can vary significantly from object to object.

This difference in the two distributions suggests that area ratio can be used to predict the presence or absence of occlusion. We can consider a simple classifier that declares an object to be unoccluded if its area ratio is greater than a threshold. For each value of the threshold, we can compute the precision and recall of this classifier. We plot the precision and recall we obtain by varying this threshold in Fig. 8b. We compute average precision and find it to be 77.17 %.



(a) (b)

Fig. 8. (a) Distribution of area ratios of occluded and unoccluded objects in the PASCAL VOC 2012 *val* set. The range of possible area ratios is discretized into 100 bins of equal width and the vertical axis shows the number object instances whose area ratios lie in a particular bin. (b) The precision-recall curve for predicting the absence of occlusion by thresholding the area ratio. (Colour figure online)

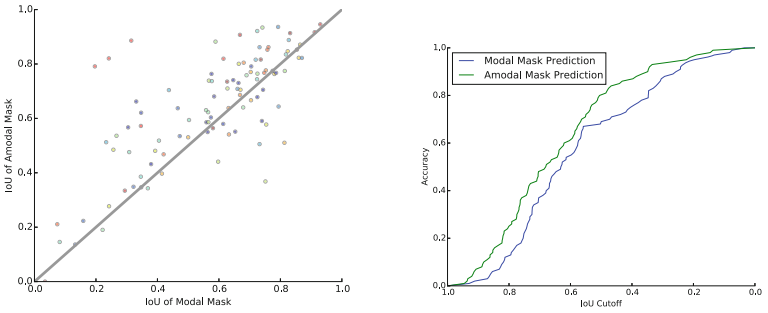
5.3 Direct Evaluation

To evaluate the accuracy of the mask produced by the proposed method, we need ground truth amodal segmentation annotations. Because no such annotations are publicly available, we collected a set of amodal segmentation masks on 100 objects. For each category, we selected five object instances randomly from the PASCAL VOC 2012 *val* set that are labelled as occluded and annotated them with amodal segmentation masks. For this purpose, we used the annotation tool for MS COCO [28], which is based on the Open Surfaces annotation tool [2].

Table 1. Performance comparison of IIS and the proposed method on the task of amodal instance segmentation

Method	Accuracy at 50 %	Accuracy at 70 %	Area under curve
IIS [27]	68.0	37.0	57.5
Proposed method	**80.0**	**48.0**	**64.3**

We compare the overlap with ground truth achieved by the amodal mask predicted by the proposed method to that achieved by the modal mask predicted by the state-of-the-art modal segmentation method, IIS. In this setting, a modal instance segmentation system represents a fairly strong baseline since in cases where occlusion is not severe, it is possible to omit the occluded portion completely from the predicted mask without significantly lowering intersection-over-union (IoU) with the ground truth.



(a) (b)

Fig. 9. (a) Comparison of overlap of the modal and amodal mask predictions with the ground truth. Overlap is measured using intersection-over-union (IoU). Each point represents an object instance and the points belonging to object instances in the same category share the same colour. Points that lie above the diagonal represent object instances whose amodal mask predictions are more accurate than their modal mask predictions. (b) Consider the setting where a predicted mask is deemed correct when it overlaps with the ground truth by at least a particular level of cutoff. This plot shows the proportion of object instances whose predicted masks are correct as a function of the cutoff. (Colour figure online)

Segmentation Performance. We first evaluate the segmentation system in isolation by taking the modal bounding box and the category of the object of interest from ground truth.

As shown in Fig. 9a, on most instances, the masks produced by the proposed method are significantly more accurate than the masks produced by IIS. Notably, the proposed method improves overlap compared to IIS by as much as 20–50 % in many cases. Overall, the proposed method produces better masks than IIS on 73 % of objects. Of the remaining 27 % of objects on which the proposed method performs worse than IIS, the drop in overlap is less than 5 % for the majority of objects. Hence, the masks produced by the proposed method are generally more accurate, sometimes by a sizeable margin.

In Fig. 9b, we plot the prediction accuracy if all masks with IoUs that exceed a particular cutoff are considered correct. As shown, predicting the mask using the proposed method consistently results in higher accuracy than using IIS at all levels of the cutoff. In Table 1, we report the accuracy of the proposed method and IIS at IoU cutoffs of 50 % and 70 %. Additionally, we compute the area under the accuracy curve for both methods. We find that the proposed method performs better than IIS on all metrics.

Combined Detection and Segmentation Performance. Next, we evaluate the performance of the combined detection and segmentation pipeline. We use faster R-CNN [32] as our detection system and compare overall performance with the proposed method as the segmentation system to IIS.

We use the amodal segmentation annotations we collected as ground truth and measure performance using mean region average precision (mAP^r) [18], which is a common metric used for modal instance segmentation. Region average precision is defined analogously to the standard average precision metric used for detection, except that overlap is computed by finding the pixel-wise IoU between the predicted and the ground truth masks. Because some instances are not annotated with ground truth amodal masks, we are unable to compute region overlap with some ground truth instances. Hence, we make a slight modification to the metric: we use bounding box overlap instead of region overlap to determine which ground truth instance a mask prediction is assigned to. However, we still use region overlap to decide if a mask prediction is deemed correct.

As shown in Table 2, the pipeline with the proposed method outperforms the pipeline with IIS by 11.1 points at 50 % overlap and 8.6 points at 70 % overlap. We also include an ablation analysis and report performance on PASCAL 3D+ [36] annotations in the supplementary material.

Table 2. Performance comparison of the combined detection and segmentation pipeline with faster R-CNN as the detection system and either IIS or the proposed method as the segmentation system

Method	mAP^r at 50 % IoU	mAP^r at 70 % IoU
Faster R-CNN [32] + IIS [27]	34.1	14.0
Faster R-CNN [32] + Proposed method	**45.2**	**22.6**

6 Conclusion

We presented a new method for amodal instance segmentation, which represents the first such method to the best of our knowledge. We introduced a novel strategy for generating synthetic amodal instance segmentation data from modal instance segmentation annotations. This strategy enabled us to train a model for amodal instance segmentation despite the lack of publicly available amodal segmentation data. Additionally, we presented a new approach for iteratively predicting the amodal bounding box from amodal segmentations. We demonstrated the effectiveness of the proposed method in predicting amodal segmentation masks both qualitatively and quantitatively.

Akcnowledgements. This work was supported by ONR MURI N00014-09-1-1051 and ONR MURI N00014-14-1-0671. Ke Li thanks the Natural Sciences and Engineering Research Council of Canada (NSERC) for fellowship support. The authors also thank Saurabh Gupta and Shubham Tulsiani for helpful suggestions and NVIDIA Corporation for the donation of GPUs used for this research.

References

1. Arbeláez, P., Hariharan, B., Gu, C., Gupta, S., Bourdev, L., Malik, J.: Semantic segmentation using regions and parts. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3378–3385. IEEE (2012)
2. Bell, S., Upchurch, P., Snavely, N., Bala, K.: OpenSurfaces: a richly annotated catalog of surface appearance. ACM Trans. Graph. (TOG) **32**(4), 111 (2013)
3. Boix, X., Gonfaus, J.M., Van de Weijer, J., Bagdanov, A.D., Serrat, J., Gonzàlez, J.: Harmony potentials. Int. J. Comput. Vis. **96**(1), 83–102 (2012)
4. Borenstein, E., Ullman, S.: Class-specific, top-down segmentation. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part II. LNCS, vol. 2351, pp. 109–122. Springer, Heidelberg (2002)
5. Carreira, J., Li, F., Sminchisescu, C.: Object recognition by sequential figure-ground ranking. Int. J. Comput. Vis. **98**(3), 243–262 (2012)
6. Dai, J., He, K., Sun, J.: Convolutional feature masking for joint object and stuff segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3992–4000 (2015)
7. Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades (2015). arXiv preprint: arXiv:1512.04412
8. Dai, Q., Hoiem, D.: Learning to localize detected objects. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3322–3329. IEEE (2012)
9. Dong, J., Chen, Q., Yan, S., Yuille, A.: Towards unified object detection and semantic segmentation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 299–314. Springer, Heidelberg (2014)
10. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1915–1929 (2013)
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
12. Fidler, S., Mottaghi, R., Yuille, A., Urtasun, R.: Bottom-up segmentation for top-down detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3294–3301 (2013)
13. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
14. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
15. Gu, C., Lim, J.J., Arbeláez, P., Malik, J.: Recognition using regions. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1030–1037. IEEE (2009)
16. Gupta, S., Arbelaez, P., Malik, J.: Perceptual organization and recognition of indoor scenes from RGB-D images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 564–571 (2013)
17. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 991–998. IEEE (2011)
18. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Simultaneous detection and segmentation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VII. LNCS, vol. 8695, pp. 297–312. Springer, Heidelberg (2014)

19. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: CVPR (2015)
20. Kanizsa, G.: Organization in Vision: Essays on Gestalt Perception. Praeger Publishers, New York (1979)
21. Kar, A., Tulsiani, S., Carreira, J., Malik, J.: Amodal completion and size constancy in natural scenes. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 127–135 (2015)
22. Kohli, P., Kumar, M.P.: Energy minimization for linear envelope MRFs. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1863–1870. IEEE (2010)
23. Kumar, M.P., Ton, P., Zisserman, A.: Obj cut. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 18–25. IEEE (2005)
24. Ladicky, L., Russell, C., Kohli, P., Torr, P.H.S.: Graph cut based inference with co-occurrence statistics. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 239–253. Springer, Heidelberg (2010)
25. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: Workshop on Statistical Learning in Computer Vision, ECCV, vol. 2, p. 7 (2004)
26. Levin, A., Weiss, Y.: Learning to combine bottom-up and top-down segmentation. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 581–594. Springer, Heidelberg (2006)
27. Li, K., Hariharan, B., Malik, J.: Iterative instance segmentation. In: CVPR (2016)
28. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 740–755. Springer, Heidelberg (2014)
29. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
30. Parkhi, O.M., Vedaldi, A., Jawahar, C., Zisserman, A.: The truth about cats and dogs. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 1427–1434. IEEE (2011)
31. Pinheiro, P., Collobert, R.: Recurrent convolutional neural networks for scene labeling. In: Proceedings of the 31st International Conference on Machine Learning (ICML 2014), pp. 82–90 (2014)
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
33. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint: arXiv:1409.1556
34. Tighe, J., Niethammer, M., Lazebnik, S.: Scene parsing with object instances and occlusion ordering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3748–3755 (2014)
35. Vijayanarasimhan, S., Grauman, K.: Efficient region search for object detection. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1401–1408. IEEE (2011)
36. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond PASCAL: a benchmark for 3D object detection in the wild. In: IEEE Winter Conference on Applications of Computer Vision, pp. 75–82. IEEE (2014)

37. Yang, Y., Hallman, S., Ramanan, D., Fowlkes, C.C.: Layered object models for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **34**(9), 1731–1743 (2012)
38. Yu, S.X., Shi, J.: Object-specific figure-ground segregation. In: Proceedings of 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. II–39. IEEE (2003)
39. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1529–1537 (2015)
40. Zhu, Y., Tian, Y., Mexatas, D., Dollár, P.: Semantic amodal segmentation (2015). arXiv preprint: arXiv:1509.01329