Amodal Segmentation through Out-of-Task and Out-of-Distribution Generalization with a Bayesian Model

Yihong Sun Adam Kortylewski Alan Yuille

Johns Hopkins University

Abstract

Amodal completion is a visual task that humans perform easily but which is difficult for computer vision algorithms. The aim is to segment those object boundaries which are occluded and hence invisible. This task is particularly challenging for deep neural networks because data is difficult to obtain and annotate. Therefore, we formulate amodal segmentation as an out-of-task and out-of-distribution generalization problem. Specifically, we replace the fully connected classifier in neural networks with a Bayesian generative model of the neural network features. The model is trained from non-occluded images using bounding box annotations and class labels only, but is applied to generalize out-of-task to object segmentation and to generalize out-of-distribution to segment occluded objects. We demonstrate how such Bayesian models can naturally generalize beyond the training task labels when they learn a prior that models the object's background context and shape. Moreover, by leveraging an outlier process, Bayesian models can further generalize out-of-distribution to segment partially occluded objects and to predict their amodal object boundaries. Our algorithm outperforms alternative methods that use the same supervision by a large margin, and even outperforms methods where annotated amodal segmentations are used during training, when the amount of occlusion is large. Code is publicly available at https://github.com/YihongSun/Bayesian-Amodal.

1. Introduction

In our everyday life, we often observe partially occluded objects. Humans can reliably recognize the visible parts of an object and use them as cues to estimate the occluded parts. This perception of the object's complete structure under occlusion is referred to as *amodal perception* [28].

In computer vision, amodal segmentation is important to study, both for its theoretical values and real-world applications. The main limitation of current approaches is the requirement of detailed supervision of amodal object masks either through human annotation [13, 24, 30] or by generating artificially occluded images [38]. Moreover, these methods assume that the object class of the occluder is known at training time. This is an important limitation in real-world applications, such as autonomous driving, where potential occluders can be any kind of real-world object.

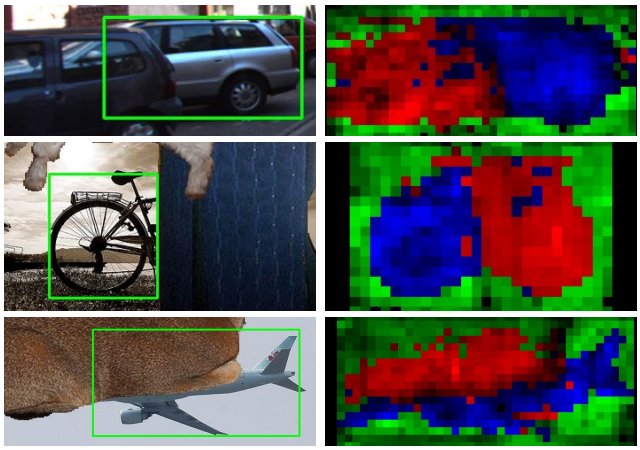


Figure 1. Our Bayesian model takes the object bounding box as input and estimates the three segmentation masks on the right: the visible object parts in blue, the invisible object parts in red, and the background context in green. The model is fully probabilistic, and the pixel brightness shows the confidence of the model prediction.

We formulate amodal segmentation as an out-of-task and out-of-distribution generalization problem, where a Bayesian generative model is trained from non-occluded objects with bounding boxes and class annotations only, but generalizes to amodal segmentation of partially occluded objects (Figure 1). Intuitively, our model can be understood as a convolutional neural network, in which the fully-connected classification head is replaced with a Bayesian generative model of the neural features. During inference, the latent model parameters (i.e. object class and amodal segmentation) are estimated such that the features of the input image are explained by the Bayesian model with max-

imum likelihood. The invariance properties of the neural features enable us to avoid explicitly modeling nuisances such as small deformations or illumination changes.

Our work builds on recent approaches of learning generative models of neural network features for image classification [19, 20], and extends these in several ways to enable amodal segmentation. In particular, we extend the network architecture with a generative model of the object's background context, as well as a prior of the object shape. Unlike standard Deep Network approaches, this makes the notion of the background context and the object shape explicit. Together, these priors enable our model to be trained from bounding box and class supervision only and generalize out-of-task to object segmentation. The Bayesian model is combined with an outlier process to make it robust to partial occlusion. The outlier process enables us to formulate amodal segmentation as an out-of-distribution task, where the model is trained from non-occluded images, but generalizes to images with partially occluded objects. We discuss how the full Bayesian model can be learned using maximum likelihood estimation with an EM-type algorithm. We also demonstrate that a joint end-to-end fine-tuning of the Bayesian model and the convolutional feature extractor further improves the performance by a steady margin.

Our experiments on all common datasets for amodal segmentation, KITTI Instance dataset (KINS) [30], COCO Amodal *cls.* [12] and Occluded-PASCAL3D+ [33], show that our Bayesian approach outperforms related weakly-supervised work by a large margin and even outperforms fully supervised methods when the amount of occlusion is large. In summary, we make several contributions:

1. We formulate **amodal instance segmentation as an out-of-task and out-of-distribution generalization problem** with a Bayesian generative model.

2. Our Bayesian model is learned from bounding box and class labels only and **outperforms alternative weakly-supervised methods by a large margin** and even outperforms supervised methods (where annotated amodal segmentations are used during training) when the amount of occlusion is large.

3. To the best of our knowledge, our model is the first for amodal segmentation that **generalizes to previously unseen occluders**.

2. Related Work

Amodal Segmentation. One of the first works in amodal segmentation is proposed by Li *et al.* [24] with an artificially generated occlusion dataset. Recently, the KINS [30] and Amodal COCO [41] datasets were introduced, which contain real-world occlusion and human estimated amodal segmentation masks. Related work on amodal segmentation follow a fully supervised approach, where either human-estimated amodal segmentation annotations are used [13, 30, 36], or synthetic occlusions are created to create training data [26, 29, 38]. However, these approaches make implicit assumptions about the amount of occlusion at test time, or even require the class of the occluder to be known at test time [29, 38]. In contrast, we introduce a Bayesian approach to this problem which is trained from non-occluded objects only and does not require any amodal supervision.

Robustness to Occlusion with Bayesian Models. Amodal segmentation is a relatively new research direction, but research on robustness to partial occlusion has received a lot of attention. In the following we focus solely on works that directly relate to ours. Recent studies [21, 40] showed that typical deep learning approaches to image classification are significantly less robust to partial occlusions than human vision. In contrast, Bayesian approaches are significantly more robust to partial occlusion, as shown in the domains of image classification [19], pose estimation [32], general object detection [20, 33], scene understanding [27, 31], face reconstruction [9] and human detection [14]. In this work, we generalize such Bayesian generative models of neural features to amodal segmentation by leveraging estimated per-pixel occlusion statistics. Notably, our work is related to Bayesian generative approaches that were developed in the pre-deep-learning-era [34]. However, our combination of modern deep learning with Bayesian generative models, enables us to generalize to very complex data with weak supervision only.

Weakly-supervised Segmentation. Due to the demanding task of acquiring expensive per-pixel annotations, many weakly-supervised instance segmentation methods have emerged that leverage cheaper labels, including image-level [1, 3, 5, 23, 39, 42] and box-level annotations [16, 17]. Notably, Zhou *et al.* [39] propose to use image-level annotations to supervise instance segmentation by exploiting class peak responses to enable a classification network for instance mask extraction. Additionally, Hsu *et al.* [16] uses box-level annotations to achieve instance segmentation by exploiting the bounding box tightness prior. Finally, Shapemask [22] addresses instance segmentation of objects with novel categories without mask annotations. Through exploiting the shape priors of known objects learned from ground truth masks, Shapemask learns the object shape and generalizes instance segmentation to novel categories. In contrast, our proposed model is able to learn shape priors without any pixel-level supervision.

3. A Bayesian Model for Amodal Segmentation

In the following, we introduce our model by first describing its input, then a simplified version of the model, and fi-

nally we proceed to develop our full Bayesian model. The structure of this section also serves to clarify the similarity and differences between our model and related work on generative models of neural network features.

3.1. The input to our model: Neural Features

Our model takes as input a feature map $\bar{F} = \psi(I, \zeta)$ at the top convolution layer of a Deep Neural Network where I is the input image and ζ are the weights of the convolutional layers. The network weights can be learnt by pre-training on ImageNet or can be directly trained end-to-end. The key property of these feature vectors is that they tend to be invariant to unimportant details of the object, which makes it easier to learn a Bayesian generative model compared to using RGB pixels as input. We denote the features within a given bounding box $\mathcal{D}$ by $F = \{f_a : a \in \mathcal{D}\}$, where a denotes the position on the lattice within the bounding box. Hence, F denotes a cropped subset of the feature map $\bar{F}$.

3.2. A Simplified Generative Model

We now discuss a simplified Bayesian generative model of the feature vectors and discuss how it can be modified to make it robust to occluders and how it can be learned. For each object, we assume that the features are generated by a mixture of distributions, which roughly correspond to the viewpoint of the objects (Equation 1). This is similar to deformable part models [10, 11] where these mixtures also have to be learnt without supervision. However, these approaches are not generative and do not address the problem of partial occlusion.

The simplest generative probability model, which corresponds to the model introduced in related work [19, 20], specifies a probability distribution:

$$P(F|y) = \sum_m p(F|y,m)P(m) = \sum_m \prod_{a\in\mathcal{D}} P_a(f_a|y,m)P(m), \tag{1}$$

$$P_a(f_a|y,m) = P_a(f_a|\mathcal{A},\Lambda) = \sum_k \alpha_{i,k}^{y,m} P(f_a|\sigma_k,\mu_k), \tag{2}$$

$$p(f|\sigma_k,\mu_k) = \frac{e^{\sigma_k \mu_k^T f}}{Z(\sigma_k)}, ||f|| = 1, ||\mu_k|| = 1, \tag{3}$$

where y denotes the object class and m refers to the mixture component. The number of mixtures is fixed a-priori and the mixture components are learnt in an unsupervised manner (similar as in Gaussian mixture models). $P(m)$ is an uniform prior over the mixture components, $\mathcal{A} = \{\alpha_{i,k}^{y,m}\}$ are mixture coefficients and $\Lambda = \{\sigma_k, \mu_k\}$ are the parameters of a von-Mises-Fisher (vMF) distribution (Equation 3). We choose a vMF distribution, because normalizing the feature vectors to unit norm makes it more feasible to estimate the model parameters in the high dimensional feature space of neural networks (note that the dimensionality of a feature vectors f_a in higher convolutional layers is typically 1024).

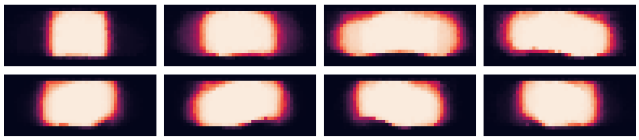


Figure 2. Compositional Shape Priors $P(\vec{w}|y,m)$. $M = 8$ compositional shape priors belonging to the car class are shown. Note that in every prior, shape and 3D pose encoding are learned by leveraging bounding box supervision only.

Learning the model parameters. In most of this paper, we assume that the parameters ζ of the Deep Network have been learnt in advance. This enables the remaining parameters of the model to be learnt by standard Bayesian methods using Maximum Likelihood via the Expectation-Maximization (EM) algorithm. Since our Bayesian model is fully differentiable, we will also discuss an alternative end-to-end learning method which learns all model parameters jointly in Section 3.5. The end-to-end training improves over the ML solution by a small but steady margin.

As shown in [19, 20], the parameters Λ correspond intuitively to a vocabulary of parts of the objects and can be learnt simply by the EM algorithm [6] initialized by the K-Means++ clustering algorithm [2]. The probability distributions $P(F|y)$ can be learnt using maximum likelihood to estimate the parameters $\mathcal{A}$. This also only requires the simple application of the EM algorithm because of the latent mixture variables m. For the sake of clarity, we refer the reader to our implementation for details on the EM learning, as the application of EM algorithm is a standard process for statistical distributions with unobserved latent variables.

Finally, the inference process is a feed-forward pass through the network to estimate $\hat{y} = \arg\max_y P(F|y)$.

Occlusion modeling. To make this model robust to occluders and enable it to generalize out-of-distribution when trained with non-occluded objects, the generative model is modified by adding an outlier process to take the form:

$$P(F|y) = \sum_m \prod_{a\in\mathcal{D}} P_a(f_a|y,m)^{z_a} Q(f_a)^{1-z_a} P(m) P(\vec{z}), \tag{4}$$

where $Q(f_a)$ is a von Mises Fisher distribution for a feature generated by an occluder estimated from unannotated images [9, 18]. The latent variable $z_a \in \{0, 1\}$ indicates whether pixel a is visible or occluded ($z_a = 1, 0$ respectively) and the prior $P(\vec{z})$ indicates the prior probability of a pixel being visible. This enables the model to not only be robust to occluders but to also simultaneously estimate the locations of the occluders $\{a \in \mathcal{D} : z_a = 0\}$ [19, 20], in addition to the object y, the mixture component m.

3.3. A Bayesian Model for Amodal Segmentation

A limitation of the simplified model described in the previous section is that it cannot segment the object, because it does not separate the foreground region corresponding to the object and a background region corresponding to the local background context of the object (e.g., the background context of an airplane will typically be sky). This motivates us to extend the model by introducing new latent variables $\{w_a\}$ to indicate foreground/background which are learnt without additional supervision. We start by extending the generative model introduced in Equation 1 to be of form:

$$P_a(f_a|y,m,w_a) = P_a(f_a|y,m)^{w_a} B_a(f_a|y,m)^{1-w_a} \times P_a(w_a|y,m) \quad (5)$$

where $w_a \in \{0,1\}$ is a latent variable indicating whether the pixel is foreground or background context ($w_a = 1, 0$ respectively). Here $\{P_a(f_a|y,m)\}$ and $\{B_a(f_a|y,m)\}$ are models for the foreground and background pixels respectively. They are specified by the foreground and background mixtures of von Mises Fisher distributions, respectively, with the same form as in Equation 2.

Shape Modeling. We introduce shape priors $P(\vec{w}|y,m) = \prod_{a\in\mathcal{D}} P_a(w_a|y,m)$, a learned 2D spatial map conditioned on the object category y and the class mixture m for the foreground/background masks as shown in Figure 2. Intuitively, they model the expected object shape for each mixture model m, and will enable the model to predict the object shape behind an occluder, as discussed in the next section. The structure of the shape priors in Figure 2 shows that the mixture components m approximately represent different 3D object poses. Finally, this gives a generative model of the data:

$$P(F|y) = \sum_{m,\vec{w}} P(F|y,m,\vec{w}) P(m) P(\vec{w}|y,m), \quad (6)$$

The model can be learned by maximizing the log-likelihood of the training data with respect to $\Lambda, \mathcal{A}, P(\vec{w}|y,m)$. This requires using the EM algorithm since the model contains latent variables for the mixtures m and the foreground/background variables $\{w_a\}$. During learning we use the standard maximum likelihood measures for vMF distributions [4], and initialize EM for the class mixtures using spectral clustering, as in [19, 20]. To initialize the foreground/background variables $\{w_a\}$, we first initialize the background distribution B_a from unannotated data (similar to estimating the distribution for the occluder Q) and initially assume that everything is foreground (i.e. $w_a = 1 \forall a$).

Occlusion Modeling. To extend this model to deal with occlusion, we also introduce an outlier process. As described in Equation 4, we introduce binary latent variables $\{z_a\}$ where $z_a \in \{0,1\}$ indicates whether pixel a is visible ($z_a = 1$) or occluded ($z_a = 0$). We introduce an occluder distribution $Q(.)$ which is a von Mises Fisher distributions $Q(f_a) = \frac{e^{\sigma \mu^T f_a}}{Z(\sigma)}, ||f_a|| = 1, ||\mu|| = 1$ whose parameters are learnt from features in unannotated images. We specify, but do not learn, a prior $P(\vec{z}) = \prod_{a\in\mathcal{D}} P(z_a)$ where $P(z=0)$ is a rough measure of how much occlusion we want the algorithm to be able to deal with.

Displacement Modeling. We also introduce a displacement variable c that models the displacement between the center of the bounding box and the center of the object. This is necessary because for partially occluded objects the bounding box only covers the visible part of the object, but amodal segmentation requires the model to predict the invisible object boundary. This gives a model of form:

$$P(F|y) = \sum_m \prod_{a\in\mathcal{D}} P_{a-c}(f_a|y,m)^{w_a z_a} \times B_{a-c}(f_a|y,m)^{(1-w_a)z_a} Q(f_a)^{(1-z_a)}. \quad (7)$$

$$P(\vec{w}|y,m,c) = \prod_{a\in\mathcal{D}} P_{a-c}(w_a|y,m) \quad (8)$$

Using this model, we can estimate the optimal object class y, class mixture m, object center c, occlusion map $\{a \in \mathcal{D} : z_a = 0\}$, and foreground map $\{a \in \mathcal{D} : w_a = 1\}$. This inference process can be implemented efficiently as a feed-forward neural network and we provide a publicly available implementation [1].

3.4. Amodal Segmentation with Our Model

After estimating distributions for the latent variables w_a and z_a, the states of w_a and z_a categorize each image pixel into one of four potential states (Figure 3). Thus, we can determine the amodal object segmentation by finding the visible and occluded foreground regions as follows.

To estimate the foreground-background segmentation w_a, we compute the posterior odds between the foreground and the background probabilities:

$$\hat{w}_a = \begin{cases} 1, & \text{if } \frac{P_{a-c}(w_a|y,m)}{(1-P_{a-c}(w_a|y,m))} \frac{P_{a-c}(f_a|y,m)}{B_{a-c}(f_a|y,m)} > 1 \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

Similarly, to infer the state of the the occlusion variables z_a, we compute the posterior odds between the occlusion and the respective foreground-background probabilities:

$$\hat{z}_a = \begin{cases} 1, & \text{if } \frac{p(z_a) P_{a-c}(f_a|y,m)^{\hat{w}_a} B_{a-c}(f_a|y,m)^{(1-\hat{w}_a)}}{(1-p(z_a))Q(f_a)} > 1 \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

Together the states of $\hat{z}_a$ and $\hat{w}_a$ allow the model to infer visible instance segmentation $\hat{M}_I = \{a : w_a = 1, z_a = 1\}$

[1] https://github.com/YihongSun/Bayesian-Amodal

Figure 3. Illustration of the four states that a pixel can be in: *Visible Foreground* ($w_a = 1, z_a = 1$) in light blue, *Occluded Foreground* ($w_a = 1, z_a = 0$) in dark blue, *Visible Background* ($w_a = 0, z_a = 1$) in black, and *Occluded Background* ($w_a = 0, z_a = 0$) in red. Consequently, amodal segmentation of an occluded object is thus defined as *Visible Foreground* $\cup$ *Occluded Foreground*.

and amodal segmentation $\hat{M}_A = \{a : w_a = 1\}$, as depicted in Figure 3. Qualitative visualization of this inference process are illustrated in Figure 1, where the relative confidence of the visible foreground $\{a : w_a = 1, z_a = 1\}$, occluded foreground $\{a : w_a = 1, z_a = 0\}$, and background $\{a : w_a = 0\}$ are represented by 3-color intensities.

3.5. End-to-End Training

When learning the parameters of the Bayesian model with the EM algorithm, we assume that the parameters of the feature extractor ζ have been initialized and fixed. This is achieved by pre-training the feature extractor for image classification using a fully-connected prediction layer and then replacing it with our Bayesian generative model. But our Bayesian model is fully differentiable and hence we can fine-tune the feature extractor and the Bayesian predictor jointly with backpropagation. This enables the feature extractor to adapt to the new predictor, which increases the models performance by a steady margin.

The objective for the end-to-end training includes a cross-entropy classification loss $\mathcal{L}_{cls}(\hat{y}, y)$ using the negative log-probability $\hat{y} = \arg\max_y -\log P(F|y)$, where $\hat{y}$ is the predicted class label and y is the true class labels. Following [19, 20], the parameters of the Bayesian model need to be trained with an additional loss ($\mathcal{L}_{ml}$) such that the Bayesian model retains a maximum likelihood of the data when the feature extractor is updated. Finally, we include an additional prior $\mathcal{L}_{seg}(\hat{M}_I, b)$ as proposed by [16] which encourages label consistency within neighboring pixels of the estimated segmentation mask. We train all parameters of our model end-to-end with γ_1 and γ_2 controlling the trade-off of the loss terms:

$$\mathcal{L} = \mathcal{L}_{cls}(\hat{y}, y) + \gamma_1 \mathcal{L}_{ml}(\Lambda, \mathcal{A}, \vec{w}) + \gamma_2 \mathcal{L}_{seg}(\hat{M}_I, b) \quad (11)$$

We note that our end-to-end trained model retains the ability to generalize out-of-distribution to partially occluded objects without ever observing partial occlusion during training. This is in contrast to standard deep networks, which do not generalize in OOD scenarios. The reason is that our model remains a generative model that is optimizing a Maximum Likelihood objective, and hence can become robust to occlusion, when equipped with an outlier process.

4. Experiments

We evaluate amodal segmentation performance of our model against a segmentation mask-supervised and a bounding box-supervised baseline on three popular amodal segmentation datasets. Due to the differences between the two baselines, we conduct experiments under two setups, one where the location of the object center is known and the other where the object center needs to be estimated.

4.1. Experimental Setup

Datasets. Following the experimental settings of related work [33], we categorize the occluded objects in each dataset into three levels of foreground occlusion from FG-1 to FG-3 and, if applicable, into three levels of background occlusion from BG-1 to BG-3.

The *OccludedVehicles* dataset [33] extends PASCAL3D+ [35] with synthetic occlusion. It contains 51801 objects evenly distributed among all occlusion levels, with both foreground and context occluded by unseen occluders.

The *KINS* dataset [30] contains real occlusion with amodal annotations. We restrict the scope of the evaluation to vehicles with a minimum height of 50 pixels, since the relevance of segmentation decreases as resolution reduces. Finally, the evaluation set contains 14826 objects.

The *COCOA-cls* dataset [12] is an extension of *Amodal COCO* [41] with class annotations, totalling 766 objects.

The *Occluded COCO* dataset [19] was introduced to test robustness of image classification to partial occlusion. It contains partially occluded objects from MS-COCO [25].

Baselines. As there is no existing model that performs amodal segmentation with class/box-level supervision only, we benchmark our model against *BBTP* [16], a state-of-the-art weakly-supervised segmentation approach, and *PCNet-M* [38], a self-supervised approach that leverages artificially generated amodal segmentation masks for training.

BBTP explores bounding box tightness prior to generate object mask under box-supervision and requires the input bounding box to be aligned to the object center c.

PCNet-M utilizes Mask RCNN [15] as an instance segmentation backbone and learns amodal completion by artificially occluding objects with other objects from the same dataset in a self-supervised manner. Hence, *PCNet-M* is considered to be the mask-supervised upper bound for our model. Since both *PCNet-M* and our model only leverage the visible parts of the object, they do not require known object center c.

Evaluation. It is observed that the occlusion levels in *KINS* are severely disproportional: over 62% of the objects

Amodal Segmentation on OccludedVehicles													
Methods	known c	superv.	FG-0	FG-1			FG-2			FG-3			Mean
			-	BG-1	BG-2	BG-3	BG-1	BG-2	BG-3	BG-1	BG-2	BG-3	
PCNet-M	✗	*mask**	**77.6**	**70.5**	**67.8**	**64.9**	**65.4**	**61.3**	**56.9**	**59.5**	**54.4**	47.6	**62.6**
Ours-ML	✗	*box*	63.3	60.2	59.9	59.8	56.9	55.6	54.8	52.6	50.2	47.1	56
Ours-E2E	✗	*box*	63	59.5	59.5	59.5	56.2	55.9	55.6	51.9	50.6	**48.3**	56
BBTP	✓	*box*	**66.5**	**59.7**	58.4	57.9	54.4	51	48.9	50.4	44.7	40.2	53.2
Ours-ML	✓	*box*	63.7	59.4	59.3	59.6	57	56.6	56.7	54.7	53.5	53.2	57.4
Ours-E2E	✓	*box*	63.9	**59.7**	**59.6**	**59.7**	**57.2**	**56.8**	**56.8**	**55**	**53.9**	**53.4**	**57.6**

Table 1. Amodal Segmentation performance evaluated on OccludedVehicles with meanIoU as the performance metric. Known c indicates whether the object center c is known and center-aligned to the proposed region. Note that 0%, 20-40%, 40-60%, and 60-80% of the object are occluded in the respective FG Occlusion Levels and 1-20%, 20-40%, and 40-60% of the context are occluded in the respective BG Occlusion Levels. Finally, PCNet-M is given additional ground truth occluder segmentation as supervision during inference, as noted by *.

Amodal Segmentation on KINS							
Methods	k. c	superv.	FG-0	FG-1	FG-2	FG-3	Mean
PCNet-M	✗	*mask*	**75.3**	65.5	52.9	33.5	56.8
Ours-ML	✗	*box*	69.2	**68.7**	62.7	45.2	61.5
Ours-E2E	✗	*box*	69.9	68.1	**63.2**	**47.3**	**62.1**
BBTP	✓	*box*	**77**	68.3	58.9	53.9	64.5
Ours-ML	✓	*box*	71.8	**70.1**	**66.2**	57.8	66.5
Ours-E2E	✓	*box*	72.3	69.6	**66.2**	**58.5**	**66.7**

Table 2. Amodal Segmentation performance is evaluated on the KINS dataset with meanIoU as the performance metric. "k.c" indicates whether the object center c is known during inference. indicates whether the object center c is known. Note that 0%, 1-30%, 30-60%, and 60-90% of the object are occluded in the respective Foreground Occlusion Levels.

Amodal Segmentation on COCOA *cls.*							
Methods	k. c	superv.	FG-0	FG-1	FG-2	FG-3	Mean
PCNet-M	✗	*mask*	56.8	53.6	47	38.4	49
Ours-ML	✗	*box*	**61.1**	**62**	**60**	**54.3**	**59.4**
Ours-E2E	✗	*box*	58.3	59.8	58.6	53.5	57.6
BBTP	✓	*box*	57.3	49.4	40.7	35	45.6
Ours-ML	✓	*box*	65	64.2	64.2	60.9	63.6
Ours-E2E	✓	*box*	**65.3**	**65**	**64.3**	**61.4**	**64**

Table 3. Transfer Evaluation from *OccludedVehicles* to *COCOA cls.* with meanIoU as the performance metric. "k.c" indicates whether the object center c is known during inference. indicates whether the object center c is known during inference. Note that 0%, 1-20%, 20-40%, and 40-70% of the object are occluded in the respective Foreground Occlusion Levels.

are non-occluded and less than 8% of objects are more than 60% occluded. Therefore, in order to examine the mask prediction quality as a function of occlusion level, we evaluate with the best region proposals (highest IoU to ground truth) generated by an RPN as supervision, removing bias towards non-occluded objects in other metrics like mAP, and separate objects into subsets based on their occlusion level.

Finally, due to the limited number of annotated objects in *COCOA cls.*, we combine the train and test set and use the combined dataset to evaluate how well models can transfer to a novel domain when trained on *OccludedVehicles*.

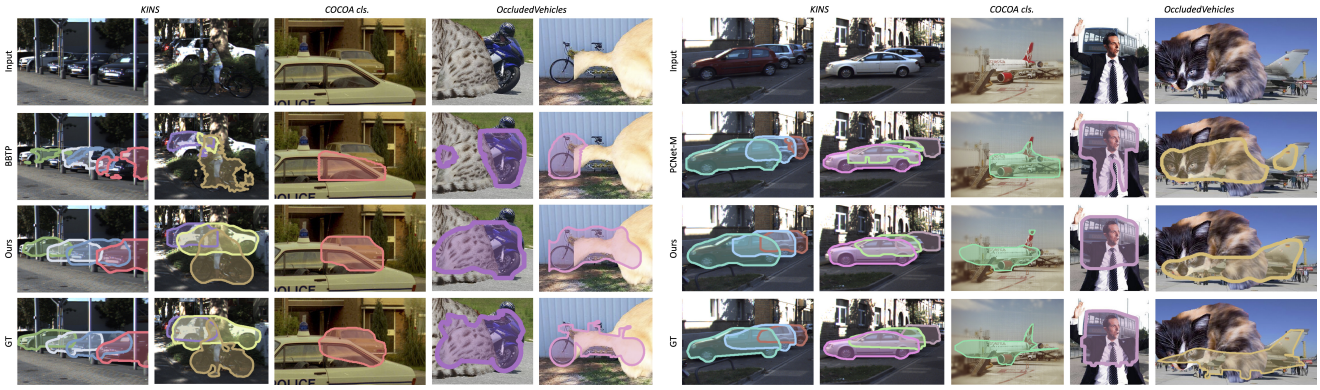
Model and Training setup. Since our Bayesian generative model is first learned with Maximum Likelihood and then fine-tuned in an end-to-end manner, we evaluate both separately, denoted as *Ours-ML* and *Ours-E2E* respectively.

Ours-ML. Our model is initially learned from feature activations ($l = 4$) of a ResNeXt-50 [37] model, pretrained on ImageNet [7]. Specifically, we initialize compositional parameters $\{\mu_k\}, \mathcal{A}, \mathcal{Z}, P(\vec{w}|y, m)$ and set the vMF variance to $\sigma_k = 65, \forall k \in \{1, \ldots, K\}$, and the number of mixtures to $M = 8$. We also learn the parameters of $n = 5$ outlier models in an unsupervised manner and fixed a prior. During initialization, we optimize the parameters with an EM algorithm as described in Section 3.3.

Ours-E2E. After learning via Maximum Likelihood, we use the obtained solution as initialization and fine-tune the model parameters as described in Section 3.5. We choose AdaGrad [8] with momentum $r = 0.98$, a learning rate of $lr = 0.01$, and trade-off weights $\gamma_1 = 2, \gamma_2 = 1$ for 10 epochs on one NVIDIA TITAN Xp for a total of 2 hours.

4.2. Results with known object center

As *BBTP* is assuming the full object bounding box (including the invisible part of the object), it is not able to estimate the amodal segmentation without knowing the object center and the corresponding full bounding box at test time. Thus, in order to evaluate and compare against *BBTP*, the object center c is given as supervision during inference,



(a) Known Object Center Comparison

(b) Unknown Object Center Comparison

Figure 4. Qualitative Amodal Segmentation Results. For cases of known and unknown ground truth object centers, we present the raw image, BBTP/PCNet-M predictions, our model's predictions, and Ground Truth from the first to fourth row, respectively.

even though our model does not necessarily require object center c to estimate the amodal segmentation.

Synthetic Occlusion. As shown on the *OccludedVehicles* dataset (Table 1), both our Bayesian model learned via Maximum Likelihood (*Ours-ML*) and fine-tuned via end-to-end training (*Ours-E2E*) outperform *BBTP* in amodal segmentation in all but two occlusion settings. Notably, in the highest occlusion level (FG-3 BG-3), our fine-tuned model is able to outperform *BBTP* by more than 13% in meanIoU.

Real Occlusion. Additionally, the trend observed in Table 1 can be confirmed by results under realistic occlusion. When evaluated on the *KINS* dataset (Table 2), both of our models outperform *BBTP* across all occlusion settings. Similarly, in the highest occlusion level, *Ours-E2E* outperforms *BBTP* by more than 4% in meanIoU.

Transferability. Seen in Table 3, our models learned via Maximum Likelihood and fine-tuned end-to-end outperform *BBTP* across all occlusion settings when learned from *OccludedVehicles* and transferred to *COCOA cls*. Notably, our end-to-end fine-tuned model outperforms *BBTP* in domain generalization and surpasses it by more than 18% in meanIoU on average. Furthermore, the observed increase in performance across all occlusion levels with known center when our model is only fine-tuned on unoccluded images further demonstrates the efficacy of the Maximum Likelihood loss term introduced in Section 3.5.

Qualitatively, shown in Figure 4 (a), it is apparent that the mask proposals generated by *BBTP* are negatively affected by the presence of occluders, while our proposed model can accurately estimate the object's amodal segmentation and preserve the object's shape consistency.

In conclusion, both quantitative and qualitative results with known object center demonstrate that our proposed model outperforms the state-of-the-art weakly-supervised method by a wide margin at amodal instance segmentation and out-of-domain transferability.

4.3. Results with unknown object center

In contrast to the previous section, since *PCNet-M* is trained with annotated occlusion, both of our models, *Ours-ML* and *Ours-E2E*, and *PCNet-M* are evaluated without the object center c given as supervision.

Synthetic Occlusion. *PCNet-M* can only perform amodal segmentation when the class label of the occluder is known a-priori in the dataset. Therefore, *PCNet-M* is inherently unsuitable to be evaluated on the *OccludedVehicles* dataset, as all occluders in the dataset belong to unseen/novel categories without explicit class annotations. Hence, in order to evaluate *PCNet-M*, we provide the *ground-truth occluder segmentation at inference time* (marked as *mask** supervised). In contrast, our approach does not require any additional information about the occluder. Seen in Table 1, even with given ground truth occluder segmentation during inference, the mask-supervised *PCNet-M* still performs worse compared to our weakly-supervised model in meanIoU at the highest occlusion level.

Real Occlusion. Furthermore, the results on the *OccludedVehicles* dataset is verified in the *KINS* dataset (Table 2), where both of our models outperform *PCNet-M* across all occluded settings. In the highest occlusion level, *Ours-E2E* outperforms *PCNet-M* by more than 13% in meanIoU.

Transferability. Similar to Section 4.2, both of our models outperform *PCNet-M* on *COCOA-cls* across all occlusion settings when transferred from *OccludedVehicles*. Notably, since the E2E model is fne-tuned with known object center and object center is unknown, our model learned from Maximum Likelihood generalizes better and surpasses *PCNet-M* by more than 10% in meanIoU on average.

Qualitatively, observed in Figure 4 (b), the mask-supervised *PCNet-M* failed to predict accurate the amodal mask of occluded objects, while our model estimates the amodal regions accurately by leveraging the prior distribu-

Shape Priors Ablation							
Methods	k. c	superv.	FG-0	FG-1	FG-2	FG-3	Mean
w/o priors	✓	*box*	61.6	59.5	58.7	58.3	59.5
w/ priors	✓	*box*	68.3	66.6	65.9	65	66.5
gt. priors	✓	*mask*	71.6	69.5	68.7	67.6	69.4

Table 4. Shape priors ablation is evaluated on the OccludedVehicles Dataset via meanIoU. Note that we report the mean performance across all BG Occlusion levels for each FG Occlusion level.

Classification on Occluded COCO							
Methods	k.c	superv.	FG-0	FG-1	FG-2	FG-3	Mean
ResNeXt-50	✓	*box*	**97.4**	85.5	81.9	56.3	80.3
CompNet	✓	*box*	94.9	89.6	84.6	**65.8**	**83.7**
CA-CompNet	✓	*box*	96	88.4	81.1	64.4	82.5
Ours-ML	✓	*box*	95	**90.4**	84	63	83.1
Ours-E2E	✓	*box*	94	89.6	**85**	**65.8**	83.6

Table 5. Classification performance evaluated on Occluded COCO. Note that 0%, 20-40%, 40-60%, and 60-80% of the object are occluded in the respective FG Occlusion Levels.

tion of the object's shapes in the mask predictions. Specifically in the right two columns of Figure 4 (b), our model is able to predict a much more realistic amodal segmentation, even though *PCNet-M* uses an additional information of a given ground-truth occluder segmentation during inference.

To conclude, our Bayesian approach outperforms the PCNet-M baseline at high occlusion levels while only requiring box-level supervision.

4.4. Ablation

In Table 4, we evaluate the effects of shape priors on amodal segmentation on the *OccludedVehicles* dataset by (1) ablating the priors (*w/o prior*), and by (2) learning the priors with ground truth segmentation (*gt. prior*). Seen in Table 4, amodal segmentation using shape priors learned from bounding box annotations significantly outperforms that without shape priors, and give comparable results as using priors learned from ground truth mask annotations.

Image Classification. Since our model uses supervision for object classification only and generalizes out-of-task to infer object segmentation, we verify the image classification performance of our model relative to related Bayesian generative models (*CompNet* [19] and *CA-CompNet* [33]) and a DCNN classifier with the same backbone under the same supervision. Seen in Table 6, our model outperforms the classifier in *BBTP* by more than 9% in classification accuracy, and outperforms the classifier in *PCNet-M* by more than 3% in classification accuracy when the object center c is unknown. Moreover, our model performs on-par with

Classification on OccludedVehicles							
Methods	k. c	superv.	FG-0	FG-1	FG-2	FG-3	Mean
PCNet-M	✗	*mask*	**98.7**	**95.9**	86.1	59.2	85
CompNet	✗	*box*	97.7	93.6	**87.3**	**73.6**	**88.1**
CA-CompNet	✗	*box*	97.7	93.4	87	73.3	87.9
Ours-ML	✗	*box*	97.7	93.4	86.8	72.6	87.6
Ours-E2E	✗	*box*	97.8	93.4	87.2	73.5	88
BBTP	✓	*box*	**99.1**	**96.6**	86	53.9	83.9
CompNet	✓	*box*	97.8	94.9	90.8	79.6	90.8
CA-CompNet	✓	*box*	98.3	95	89.7	76.6	89.9
Ours-ML	✓	*box*	97.8	95.2	90.7	80.2	91
Ours-E2E	✓	*box*	98.3	95.6	**91.4**	**81.4**	**91.7**

Table 6. Classification performance evaluated on OccludedVehicles. Note that a mean is taken across all BG Occlusion levels.

CompNet and CA-CompNet. Similarly, found in Table 5, our model outperforms ResNeXt-50 by more than 3 % when evaluated on Occluded COCO with real occlusions.

In summary, our model performs favorably over BBTP and PCNet-M in terms of image classification. It also performs on par with CompNets but can additionally perform amodal perception reliably, while being trained from bounding box and class-level supervision only.

5. Conclusion

In this work, we studied the problem of amodal segmentation from the perspective of out-of-task and out-of-distribution generalization with a Bayesian model. We learn a Bayesian generative model of neural network features, which explicitly represents the object's background context and foreground shape. This enables the model to localize occluded object parts and predict the occluded object shape. Our Bayesian approach for amodal segmentation only requires bounding box and class supervision, achieving state-of-the-art performance at amodal segmentation when compared to other weakly-supervised method and even outperforming fully supervised methods at high occlusion levels.

Limitations and Societal Impact. One limitation of our work is the dependence on 2D shape priors, which would require a large number to properly represent highly non-rigid objects such as humans or animals. For future work, we therefore expect that the learning of 3D shape priors would render the model more efficient and would also enhance the generalization ability to previously unseen 3D poses. Like most segmentation works, our work does not introduce any foreseeable societal impacts, but will generally promote more data-efficient and robust computer vision models.

Acknowledgements. We gratefully acknowledge funding support from Office of Naval Research (N00014-21-1-2812) and National Science Foundation (BCS-1827427).

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019. 2

[2] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006. 3

[3] Aditya Arun, CV Jawahar, and M Pawan Kumar. Weakly supervised instance segmentation by learning annotation consistent instances. *arXiv preprint arXiv:2007.09397*, 2020. 2

[4] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(9), 2005. 4

[5] Hisham Cholakkal, Guolei Sun, Fahad Shahbaz Khan, and Ling Shao. Object counting and instance segmentation with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 12397–12405, 2019. 2

[6] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. 3

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 6

[8] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011. 6

[9] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision*, 126(12):1269–1287, 2018. 2, 3

[10] Pedro F Felzenszwalb, Ross B Girshick, and David McAllester. Cascade object detection with deformable part models. In *2010 IEEE Computer society conference on computer vision and pattern recognition*, pages 2241–2248. Ieee, 2010. 3

[11] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International journal of computer vision*, 61(1):55–79, 2005. 3

[12] Patrick Follmann, Rebecca König, Philipp Härtinger, Michael Klostermann, and Tobias Böttger. 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, 5

[13] 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. 1, 2

[14] Ross Girshick, Pedro Felzenszwalb, and David McAllester. Object detection with grammar models. *Advances in neural information processing systems*, 24, 2011. 2

[15] 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. 5

[16] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6586–6597. Curran Associates, Inc., 2019. 2, 5

[17] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 2

[18] Adam Kortylewski. *Model-based image analysis for forensic shoe print recognition*. PhD thesis, University_of_Basel, 2017. 3

[19] Adam Kortylewski, Ju He, Qing Liu, and Alan L Yuille. Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8940–8949, 2020. 2, 3, 4, 5, 8

[20] Adam Kortylewski, Qing Liu, Angtian Wang, Yihong Sun, and Alan Yuille. Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion. *International Journal of Computer Vision*, 2020. 2, 3, 4, 5

[21] Adam Kortylewski, Qing Liu, Huiyu Wang, Zhishuai Zhang, and Alan Yuille. Combining compositional models and deep networks for robust object classification under occlusion. *The IEEE Winter Conference on Applications of Computer Vision*, March 2020. 2

[22] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. Shapemask: Learning to segment novel objects by refining shape priors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9207–9216, 2019. 2

[23] Issam H Laradji, David Vazquez, and Mark Schmidt. Where are the masks: Instance segmentation with image-level supervision. *arXiv preprint arXiv:1907.01430*, 2019. 2

[24] Ke Li and Jitendra Malik. Amodal instance segmentation. In *European Conference on Computer Vision*, pages 677–693. Springer, 2016. 1, 2

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[26] Huan Ling, David Acuna, Karsten Kreis, Seung Wook Kim, and Sanja Fidler. Variational amodal object comple-

tion. *Advances in Neural Information Processing Systems*, 33:16246–16257, 2020. 2

[27] Pol Moreno, Christopher KI Williams, Charlie Nash, and Pushmeet Kohli. Overcoming occlusion with inverse graphics. In *European Conference on Computer Vision*, pages 170–185. Springer, 2016. 2

[28] Bence Nanay. The importance of amodal completion in everyday perception. *i-Perception*, 9(4):2041669518788887, 2018. 1

[29] Khoi Nguyen and Sinisa Todorovic. A weakly supervised amodal segmenter with boundary uncertainty estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7396–7405, 2021. 2

[30] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 5

[31] Lukasz Romaszko, Christopher KI Williams, Pol Moreno, and Pushmeet Kohli. Vision-as-inverse-graphics: Obtaining a rich 3d explanation of a scene from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 851–859, 2017. 2

[32] Angtian Wang, Adam Kortylewski, and Alan Yuille. Nemo: Neural mesh models of contrastive features for robust 3d pose estimation. In *International Conference on Learning Representations*, 2021. 2

[33] Angtian Wang, Yihong Sun, Adam Kortylewski, and Alan L Yuille. Robust object detection under occlusion with context-aware compositionalnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12645–12654, 2020. 2, 5, 8

[34] Christopher KI Williams and Michalis K Titsias. Greedy learning of multiple objects in images using robust statistics and factorial learning. *Neural Computation*, 16(5):1039–1062, 2004. 2

[35] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, pages 75–82. IEEE, 2014. 5

[36] Yuting Xiao, Yanyu Xu, Ziming Zhong, Weixin Luo, Jiawei Li, and Shenghua Gao. Amodal segmentation based on visible region segmentation and shape prior. *arXiv preprint arXiv:2012.05598*, 2020. 2

[37] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 6

[38] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 5

[39] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3791–3800, 2018. 2

[40] Hongru Zhu, Peng Tang, Jeongho Park, Soojin Park, and Alan Yuille. Robustness of object recognition under extreme occlusion in humans and computational models. *CogSci Conference*, 2019. 2

[41] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollar. Semantic amodal segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 5

[42] Yi Zhu, Yanzhao Zhou, Huijuan Xu, Qixiang Ye, David Doermann, and Jianbin Jiao. Learning instance activation maps for weakly supervised instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3116–3125, 2019. 2