

# Robust Instance Segmentation through Reasoning about Multi-Object Occlusion

Xiaoding Yuan<sup>1</sup> Adam Kortylewski<sup>2</sup> Yihong Sun<sup>2</sup> Alan Yuille<sup>2</sup>

<sup>1</sup>Tongji University, <sup>2</sup>Johns Hopkins University

## Abstract

Analyzing complex scenes with Deep Neural Networks is a challenging task, particularly when images contain multiple objects that partially occlude each other. Existing approaches to image analysis mostly process objects independently and do not take into account the relative occlusion of nearby objects. In this paper, we propose a deep network for multi-object instance segmentation that is robust to occlusion and can be trained from bounding box supervision only. Our work builds on Compositional Networks, which learn a generative model of neural feature activations to locate occluders and to classify objects based on their non-occluded parts. We extend their generative model to include multiple objects and introduce a framework for efficient inference in challenging occlusion scenarios. In particular, we obtain feed-forward predictions of the object classes and their instance and occluder segmentations. We introduce an Occlusion Reasoning Module (ORM) that locates erroneous segmentations and estimates the occlusion order to correct them. The improved segmentation masks are, in turn, integrated into the network in a top-down manner to improve the image classification. Our experiments on the KITTI INSTance dataset (KINS) and a synthetic occlusion dataset demonstrate the effectiveness and robustness of our model at multi-object instance segmentation under occlusion. Code is publically available at <https://github.com/XD7479/Multi-Object-Occlusion>.

## 1. Introduction

Scenes in images most often depict multiple objects that partially occlude each other. Recent studies [38, 18] showed that deep networks are less robust at recognizing partially occluded objects compared to Humans. The main difficulties are raised by the combinatorial variability of the object ordering and positioning, as well as the fact that scenes can contain known and unknown object classes.

One approach to address the problem of occlusion in

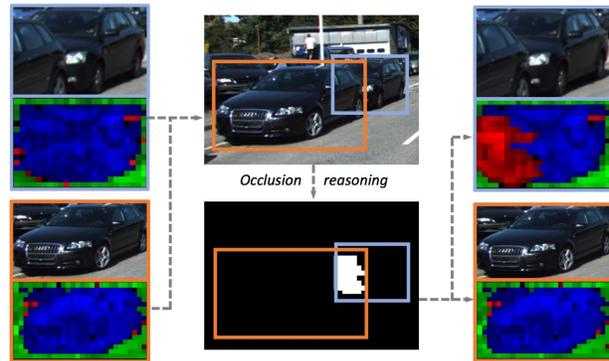


Figure 1: Our proposed model corrects erroneous instance segmentations through multi-object reasoning. Left: Two input images that are processed independently. The segmentation results identify visible object parts in blue, invisible parts in red, and context in green. Note how in the top image, the model cannot identify the occlusion. Center: By enforcing consistency between segmentations of nearby objects, our model can identify conflicting segmentations (white area). Right: Reasoning about the occlusion order resolves the erroneous predictions.

deep networks is data augmentation [36, 5, 34, 1]. While this increases the robustness of deep networks, the classification performance on partially occluded objects still remains substantially worse compared to non-occluded objects. Recent work introduced compositional deep networks (CompositionalNets) and showed that these are more robust to partial occlusion compared to data augmentation approaches [16, 17, 30]. CompositionalNets are deep neural network architectures in which the fully connected classification head is replaced with a differentiable compositional model. The structure of the compositional model enables CompositionalNets to decompose images into objects and context, as well as to further decompose objects into their individual parts. The generative nature of the compositional model enables it to segment objects and occluders

[28] and to recognize objects based on their non-occluded parts. However, CompositionalNets, as well as other popular architectures, treat each object in an image independently and do not explicitly exploit the mutual relationship of nearby objects.

In this paper, we introduce a deep network for multi-object instance segmentation that is robust to occlusion and can be trained from bounding box supervision only. Our work builds on and significantly extends CompositionalNets. Specifically, we extend the generative model in CompositionalNets to allow for instance segmentation of multiple mutually occluding objects in an image. This multi-object generative model is hard to optimize because of the mutual dependencies between objects. To solve this optimization efficiently, we introduce an Occlusion Reasoning Module (ORM) that takes as input the independent predictions of each objects label, the instance segmentation and the occluder segmentation (Figure 1). We proceed to estimate possibly erroneous predictions through an occlusion voting mechanism. During occlusion voting, each object in the image votes for every pixel in its bounding box if the pixel is occupied by the object or if is occluded. Pixels which receive ambiguous votes from multiple objects indicate segmentation errors. To correct these we leverage the occlusion order of overlapping bounding boxes based on the classification scores. The corrected instance and occlusion segmentation masks are fed back into the CompositionalNet to mask out those features that induced segmentation errors, and to improve the prediction of the object class.

We perform extensive experiments on the KITTI Instance dataset (KINS). We further introduce a synthetic dataset that comprises artificially generated images of partially occluded objects, which are generated by superimposing segmented objects from the KITTI. The synthetic generation of partially occluded images enables us to evaluate custom types of occlusion challenges such as: pair-wise occlusion, multi-object occlusion and mixed occlusion containing both known and unknown object classes as occluders. Our experimental results highlight that reasoning about multi-object occlusion significantly enhances the robustness of deep networks as it enables them to detect erroneous feed-forward predictions and self-correct through reasoning about multi-object occlusion. In summary, our contributions in this work are:

1. We introduce a deep network for multi-object instance segmentation that is robust to occlusion and can be trained from bounding box supervision only. Specifically, our network defines a generative model of multiple objects and achieves enhanced robustness through reasoning about multi-object occlusion.

2. We introduce an Occlusion Reasoning Module (ORM) that enables efficient inference in generative models with multiple objects. In particular, it detects erroneous

feed-forward predictions and corrects them through reasoning about the occlusion order of objects.

3. We achieve state-of-the-art performance at instance segmentation under occlusion on the KITTI Instance (KINS) dataset.

4. We introduce an occlusion challenge generated from real-world segmented objects with accurate annotations and propose a taxonomy of occlusion scenarios that pose a particular challenge for computer vision.

## 2. Related Work

**Occlusion reasoning.** A number of approaches have recently been proposed to integrate occlusion reasoning in areas including image classification [17, 32], object detection [30], segmentation [8, 31] and tracking [33]. Gao *et al.*[8] introduce binary variables to infer the visible cells in a bounding box. Hsiao and Hebert [13] model occlusions by reasoning about 3D relationship of objects approximated by their bounding boxes. Recent works on pixel-level occlusion reasoning include a probabilistic model proposed by George *et al.*[10] that contains mutual occlusion inference on text-based CAPTCHAs by approximating MAP solution through message passing. Another probabilistic framework by Yanget *et al.*[33] introduce occlusion priori modeled by Markov random field to tackle mutual occlusion in object tracking task. Tighe *et al.*[29] introduce an inter-class occlusion prior to parse scenes and refine pixel-level labels. OFNet designed by Lu *et al.*[22] considers the relevance between occlusion contours and pixel orientations, but no semantic information is included. Zhan *et al.*[35] propose pair-wise order recovery by comparing the amodal mask completion of neighboring objects in a self-supervised way, while lack the ability of handling unknown occlusion. Our proposed architecture performs pixel-level occlusion reasoning and ensures the consistence of object shape by object-level occlusion order recovery. Note that we can handle both the unknown occlusion and multi-object occlusion at the same time.

**Weakly-supervised instance segmentation.** While instance segmentation performance was significantly advanced by CNN based architectures [11, 3, 21, 2], pixel-level semantic annotation is required for training by fully supervised methods. Weakly-supervised segmentation methods require only image-level supervision [27, 23] and bounding-box-level annotations [25, 20] to reduce the cost of dense labeling. DeepCut proposed by Rajchl *et al.* [25] extends GrabCut [26] by training a CNN as classifier from bounding box annotations and address instance segmentation as energy minimisation problem based on conditional random fields. Zhou *et al.*[37] present an instance mask extraction by class response maps indicating visual cues with

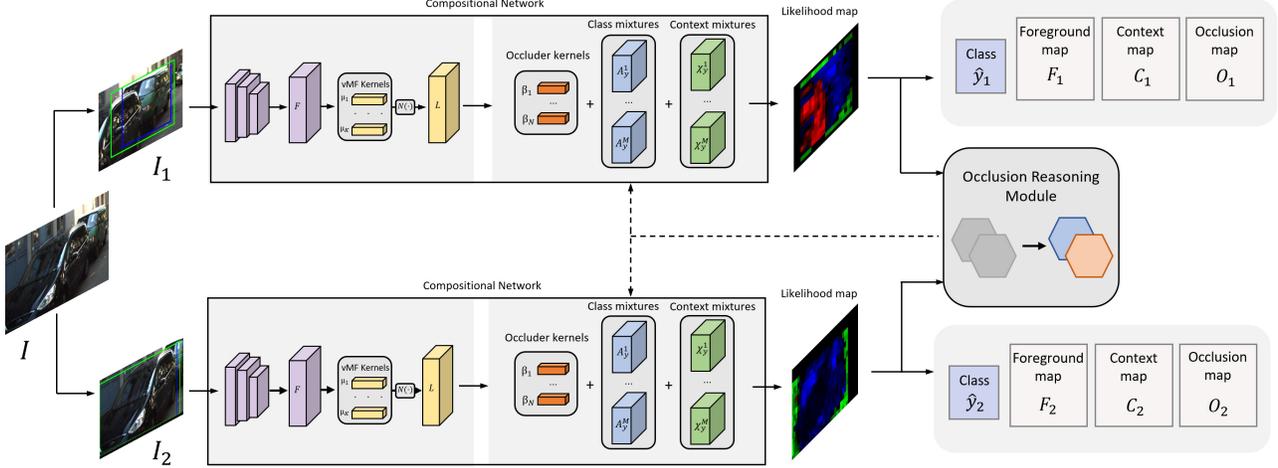


Figure 2: Proposed deep network architecture for multi-object instance segmentation under occlusion. Given an input image, we crop objects based on their bounding box. Each crop is processed by a Compositional Network to obtain independent estimates of the object class, instance segmentation and occlusion segmentation. Subsequently, these are processed by the multi-object reasoning module, which detects inconsistent segmentations and corrects them by taking into account the occlusion order of the objects. The corrected instance segmentation mask is used in a top-down manner to mask out occluded features, which, in turn, improves the classification score. Note we draw two Compositional Networks for illustrative purpose, in practice the images are processed sequentially by the same network.

image-level supervision. Hsu *et al.*[14] address the problem as multiple instance learning task and estimate the foreground/background by generating positive/negative bags based on the sweeping lines of each bounding box. Amodal instance segmentation task were introduced more recently. Li *et al.*[19] firstly presented a solution for amodal instance segmentation training with artificial occlusion. Other methods [39, 24, 7] implement fully-supervised amodal mask completion. In this work, we build on the weakly-supervised instance segmentation CompositionalNets[28], and generalize them to allow for reasoning about multi-object occlusion.

### 3. Robustness through Occlusion Reasoning

**Notation.** The output of the layer  $l$  in a DCNN is referred to as *feature map*  $F^l = \psi(I, \Omega) \in \mathbb{R}^{H \times W \times D}$ , where  $I$  and  $\Omega$  are the input image and the parameters of the feature extractor, respectively. *Feature vectors* are vectors in the feature map,  $f_i^l \in \mathbb{R}^D$  at position  $i$ , where  $i$  is defined on the 2D lattice of  $F^l$  with  $D$  being the number of channels in the layer  $l$ . We omit subscript  $l$  in the following for clarity since the layer  $l$  is fixed a priori in the experiments.

#### 3.1. Prior Work: CompNets for Single Objects

CompositionalNets [16, 17] are deep neural network architectures in which the fully connected classification head is replaced with a differentiable compositional model. In particular, the classification head defines a generative model

$p(F|y)$  of the features  $F$  for an object class  $y$ :

$$p(F|\Theta_y) = \sum_m \nu_m p(F|\theta_y^m), \nu_m \in \{0, 1\}, \sum_{m=1}^M \nu_m = 1 \quad (1)$$

Here  $M$  is the number of mixtures of compositional models per each object category and  $\nu_m$  is a binary assignment variable that indicates which mixture component is active.  $\Theta_y = \{\theta_y^m = \{\mathcal{A}_y^m, \chi_{i,y}^m, \Lambda\} | m=1, \dots, M\}$  are the overall compositional model parameters for the category  $y$ . The individual mixture components are defined as:

$$p(F|\theta_y^m) = \prod_i p(f_i | \mathcal{A}_{i,y}^m, \chi_{i,y}^m, \Lambda) \quad (2)$$

Note how the distribution *decomposes* the feature map  $F$  into a set of individual feature vectors  $f_i$ .  $\mathcal{A}_y^m = \{\mathcal{A}_{i,y}^m | i \in [H, W]\}$  and  $\chi_y^m = \{\chi_{i,y}^m | i \in [H, W]\}$  are the parameters of the mixture components.

The feature likelihood is defined as composition of a foreground and a context likelihood:

$$p(f_i | \mathcal{A}_{i,y}^m, \chi_{i,y}^m, \Lambda) = p(i|m, y) p(f_i | \mathcal{A}_{i,y}^m, \Lambda) \quad (3)$$

$$+ (1 - p(i|m, y)) p(f_i | \chi_{i,y}^m, \Lambda). \quad (4)$$

The parameters of the foreground and context likelihood are  $\mathcal{A}_{i,y}^m$  and  $\chi_{i,y}^m$  respectively.  $p(i|m, y)$  is a prior that models how likely a feature vector at position  $i$  is to be located in the foreground. In particular,  $\mathcal{A}_{i,y}^m = \{\alpha_{i,k,y}^m | k =$

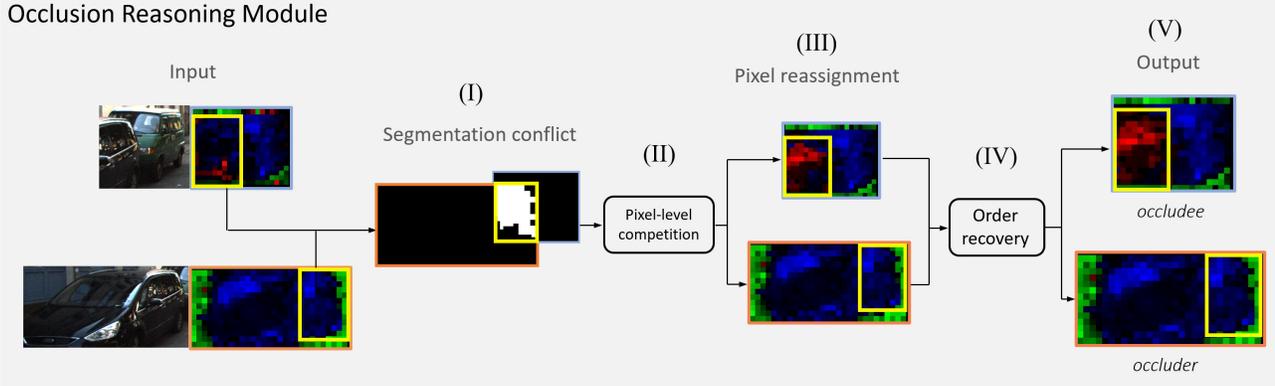


Figure 3: Detailed method in Occlusion Reasoning Module (ORM). The input of ORM is the segmentation likelihood maps of a pair of neighboring objects. The map contains pixel-level prediction into foreground (blue pixels), background (green pixels) and occlusion (red pixels). Brighter pixel refers to higher likelihood on each position. Segmentation conflict is detected when one pixel is defined as foreground for both objects. Pixel-level competition is performed to solve the conflict and re-assign pixels. Pair-wise occlusion order recovery results from the competition, and the occludee’s likelihood map is then updated accordingly.

$1, \dots, K$  are mixture coefficients and  $\Lambda = \{\lambda_k = \{\sigma_k, \mu_k\} | k = 1, \dots, K\}$  are the parameters of von-Mises-Fisher distributions:

$$p(f_i | \mathcal{A}_{i,y}^m, \Lambda) = \sum_k \alpha_{i,k,y}^m p(f_i | \lambda_k), \quad (5)$$

$$p(f_i | \lambda_k) = \frac{e^{\sigma_k \mu_k^T f_i}}{Z(\sigma_k)}, \quad \|f_i\| = 1, \|\mu_k\| = 1. \quad (6)$$

The context likelihood  $p(f_i | \chi_{i,y}^m, \Lambda)$  is defined accordingly. Note that  $K$  is the number of components in the vMF mixture distributions and  $\sum_{k=0}^K \alpha_{i,k,y}^m = 1$ .  $Z(\sigma_k)$  is the normalization constant. The priors  $p(i|m, y)$  and likelihood parameters can be learned by segmenting the training images into foreground and background. We follow the approach introduced in [30] which uses weakly supervised segmentation based on the bounding box annotation to segment the context from the object. All model parameters  $\{\Omega, \{\Theta_y\}\}$  can be trained end-to-end as discussed in [16, 28].

**Partial Occlusion.** Compositional networks can be augmented with an outlier model to enhance their robustness to partial occlusion. The intuition is that at each position  $i$  in the image either the object model  $p(f_i | \mathcal{A}_{i,y}^m, \chi_{i,y}^m, \Lambda)$  or an outlier model  $p(f_i | \beta, \Lambda)$  is active:

$$p(F | \theta_y^m, \beta) = \prod_i p(f_i | \beta, \Lambda)^{1-z_i^m} p(f_i | \mathcal{A}_{i,y}^m, \chi_{i,y}^m, \Lambda)^{z_i^m}. \quad (7)$$

The binary variables  $\mathcal{Z}^m = \{z_i^m \in \{0, 1\} | i \in \mathcal{P}\}$  indicate if the object is occluded at position  $i$  for mixture component  $m$ .

The outlier model is defined as:

$$p(f_i | \beta, \Lambda) = \sum_k \beta_{n,k} p(f_i | \sigma_k, \mu_k). \quad (8)$$

Note that the model parameters  $\beta$  are independent of the position  $i$  in the feature map and thus the model has no spatial structure. The parameters of the occluder models  $\beta$  are learned from clustered features of random natural images that do not contain any object of interest [16].

**Instance segmentation with CompositionalNets.** Sun et al.[28] showed that instance segmentation can be achieved with CompositionalNets by simply comparing the likelihood terms of the model. In particular, we can predict the pixel-wise labels to be foreground  $\mathcal{F}$ , context  $\mathcal{C}$  or occlusion  $\mathcal{O}$  by computing the respective likelihoods:

$$p(f_i = \mathcal{O}) = p(i|m, y) p(f_i | \beta, \Lambda) \quad (9)$$

$$p(f_i = \mathcal{F}, y) = p(i|m, y) p(f_i | \mathcal{A}_{i,y}^m, \Lambda) \quad (10)$$

$$p(f_i = \mathcal{C}, y) = (1 - p(i|m, y)) p(f_i | \chi_{i,y}^m, \Lambda) \quad (11)$$

### 3.2. Compositional Networks for Multiple Objects

The main limitation of Compositional Networks is that they assume only one object is present in an image. They can be trivially generalized to multiple objects by treating each object independently [16, 30]. However, assuming independence between objects neglects the relations between them and leads to inconsistencies in the segmentation results. For example Figure 3 shows how two objects with overlapping bounding boxes both predict that they are visible in the overlapped region. Whereas, it is clear that only one object can be visible per pixel in an image.

In this work, we aim to resolve such inconsistencies by enabling deep networks to reason about multi-object occlusion. In particular, we generalize the generative model in compositional networks to multiple objects by extending

the model likelihood:

$$p(F|\theta_{y_1}^m, \dots, \theta_{y_N}^m, \beta) = \prod_i \prod_{n=1}^{N+1} p_n(f_i)^{z_{i,n}}, \quad (12)$$

with  $\sum_n z_{i,n}=1$  and  $z_{i,n} \in \{0, 1\}$ . This generalized likelihood includes  $n=\{1, \dots, N\}$  object models  $p_n(f_i)=p(F|\theta_{y_n}^m, \beta)$ , which correspond to the number of objects in the image, and the outlier model  $p_{N+1}(f_i)=p(f_i|\beta, \Lambda)$ . Note that, by the design of the likelihood, only one object model can be active at any location  $i$  in the feature map  $F$ . Maximizing the model likelihood defined in Equation 12 is difficult because it involves multiple objects and the visibility at each pixel  $z_{i,n}$  depends on the visibility of the neighboring pixels. We solve this complex optimization problem by introducing a multi-object reasoning module into the architecture of CompositionalNets.

### 3.3. Reasoning about Multi-Object Occlusion

In this section, we introduce a deep network for multi-object instance segmentation that is robust to occlusion and can be trained from bounding box supervision only. To make our discussion concise, we constrain ourselves in this section to images that contain two objects, where one partially occludes the other. Note, however, that our model trivially extends to multiple objects.

**Feed-forward extraction of likelihood maps.** Our proposed network architecture is illustrated in Figure 2. We draw two CompositionalNet architectures to enhance the clarity of the illustration, in practice they are sequentially processed by the same network. The objects in the input image  $I$  are cropped based on their bounding box and first independently processed by a CompositionalNet. For each image crop  $I_1, I_2$  we obtain a class prediction  $\hat{y}_1, \hat{y}_2$  and three likelihood maps that encode the foreground, context and occlusion likelihood in the feature map:

$$\mathcal{F}_1 = \{p(f_i = \mathcal{F}, \hat{y}_1) | \forall i \in \mathcal{P}\} \quad (13)$$

$$\mathcal{C}_1 = \{p(f_i = \mathcal{C}, \hat{y}_1) | \forall i \in \mathcal{P}\} \quad (14)$$

$$\mathcal{O}_1 = \{p(f_i = \mathcal{O}) | \forall i \in \mathcal{P}\} \quad (15)$$

we compute  $\mathcal{F}_2, \mathcal{C}_2, \mathcal{O}_2$  respectively. We illustrate the likelihood maps throughout the paper as a two-dimensional heat map that is color coded. In particular, we visualize at each pixel which likelihood has the highest value by coloring occluder red, foreground blue and context green. The pixel intensity encodes the difference between the three likelihood terms. For dark pixels all likelihoods have similar values, hence indicating that the model is uncertain, whereas at bright pixels one likelihood is clearly higher compared to the other. As shown in Figure 3, instance segmentation based on the likelihood maps can be incorrect in the region

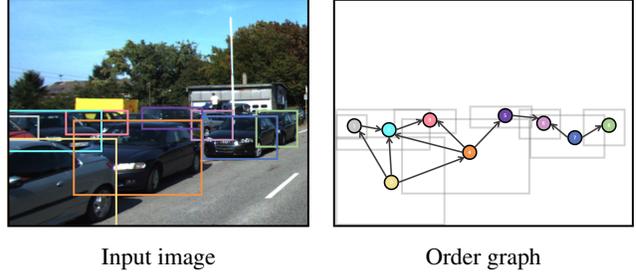


Figure 4: Occlusion order graph recovered by the proposed network. Left: Input image with bounding boxes; Right: Occlusion order graph, where the direction of the arrows indicates occlusion. Note how the correct ordering is recovered in a very challenging occlusion scenario.

where the two bounding boxes overlap (yellow box), particularly, for the occluded object. In practice, we observe that errors occur most often when an object is occluded by another object of the same category. As discussed earlier, this is caused by the fact that the segmentation is performed independently of other objects in an image, and in addition also per pixel independently. While these independence assumptions enable an efficient feed-forward inference, they neglect important relationships in images. For example in the overlapping region of the bounding boxes in Figure 3, we want all pixels to be assigned to the same object. It is very unnatural to treat every pixel independently. We propose a multi-object reasoning module that resolves such segmentation conflicts by taking into account additional relationships between objects at minimal computational overhead.

**Pixel-level competition.** Figure 3 illustrates the pipeline of the occlusion reasoning module. We first detect segmentation conflicts as those image pixels are classified as foreground by both object models (Figure 3, I). We denote this conflict set as  $\mathcal{C}$ . Note that for the occludee (the occluded object), some of the feature vectors in the occluded region are mis-classified as foreground, however, their likelihoods are lower compared to those of the occluder at the same pixel (indicated by the intensity of the color).

We exploit this by taking into account the relationship between both objects. In particular, we assign the feature vector  $f_i$  to one of the two objects by comparing their foreground likelihoods (Figure 3, II):

$$z_{i,1} = \begin{cases} 1, & \text{if } p(f_i = \mathcal{F}, \hat{y}_1) > \max\{p(f_i = \mathcal{F}, \hat{y}_2), p(f_i = \mathcal{O})\} \\ 0, & \text{otherwise} \end{cases} \quad (16)$$

We compute the visibility variables of the second object  $z_{i,2}$  and the outlier model  $z_{i,3}$  accordingly. Using the estimated visibility variables, we can re-assign each pixel in the corresponding segmentation maps (Figure 3, III).

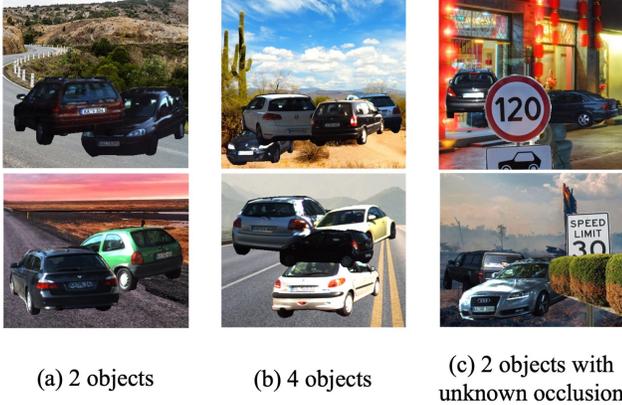


Figure 5: We synthesis three challenging occlusion scenarios: Occlusion with (a) 2 objects; (b) 4 objects with complex inter-object occlusion; and (c) multi-object occlusion including objects with of previously unseen occlusion.

**Order recovery.** At this stage, each pixel has been updated independently. However, it is natural to assume that in the overlapping region of two bounding boxes only one object is the foreground object, whereas the other object is in the background. Hence all pixels should be assigned to either of the two objects. To encode this property we estimate the occlusion order between the objects (Figure 3, IV). Specifically, we estimate the occlusion order  $R(I_1, I_2)$  by comparing the number of pixels assigned to each object in the region of the segmentation conflict:

$$R(I_1, I_2) = \begin{cases} 1, & \sum_{i \in C} z_{i,1} > \sum_{i \in C} z_{i,2} \\ -1, & \text{otherwise} \end{cases} \quad (17)$$

Figure 4 illustrates the effectiveness of this approach at recovering the occlusion order, even in challenging, multi-object occlusion scenarios. Using the predicted occlusion order  $R(I_1, I_2)$ , we reassign the visibility variables  $z_{i,n}$  in all "all or nothing" manner, such that all the variables that are not assigned to the outlier model are assigned to the object in the front. Figure 3 illustrates how the multi-object occlusion reasoning benefits the instance segmentation compared to the input segmentation which was achieved by processing the images independently.

**Self-correction through occlusion updates.** Compared to the occlusion variables  $Z^m$ , which were estimated in the feed-forward stage, the newly estimated visibility variables  $z_{i,n}$  take into account the knowledge of neighboring objects and their occlusion order graph. This newly acquired knowledge can subsequently be used to recompute the model likelihood  $p(F|\Theta_y)$  of the occluded object, by replacing the occlusion variables in Equation 7. As our experimental results demonstrate, this top-down refinement enables CompositionalNets to correct miss-classifications that were induced by wrongly estimated occlusion vari-

	Mask	L0	L1	L2	L3	Mean
Mask R-CNN	✓	85.8	81.5	72.7	51.9	73
CompNet	✗	75.8	67.7	44.4	23.3	64.3
Ours (iter=2)	✗	<b>75.9</b>	<b>69.2</b>	<b>54.0</b>	<b>34.6</b>	<b>67.2</b>

	Mask	L0	L1	L2	L3	Mean
PCNet-M	✓	83.1	77.5	68.5	51.6	70.2
BBTP	✗	77.9	71.6	67	67.8	71.1
CompNet	✗	76.6	76.1	75.9	74.7	76.2
Ours (iter=2)	✗	<b>76.9</b>	<b>76.4</b>	<b>76.5</b>	<b>76.5</b>	<b>76.7</b>

Table 1: Modal and amodal instance segmentation on the KINS dataset (top and bottom). We compare to fully-supervised Mask R-CNN, self-supervised PCNet-M, weakly-supervised BBTP, and CompNets with and without ORM. Occlusion levels L1-L3 are defined as: L1: 1%-30%, L2: 30%-60%, L3: 60%-90% of the object is occluded.

ables. This will particularly improve the classification performance of occluded objects by a large margin. We repeat the self-correction through multi-object reasoning recurrently as the updated classification score can lead to changes in the assignment of the mixture models, and hence can lead to improved segmentations.

## 4. Experiments

We evaluate our deep network for multi-instance segmentation under occlusion on the KINS dataset and on an artificially generated occlusion challenge dataset. We will present experimental results of weakly-supervised modal and amodal instance segmentation, and ablate the occlusion reasoning module for order recovery.

### 4.1. Datasets

**KINS.** The KINS dataset [24] is augmented from KITTI [9] with more instance pixel-level annotation for 8 categories including amodal instance segmentation and relative occlusion order. Amodal instance segmentation aims at segmenting the complete instance shape, even when the object is only partially visible. The dataset contains 7474 images for training and 7517 for testing.

**Occlusion Challenge.** The amodal segmentation predicted on 2D real-world images by human judgements is still subjective and imprecise. Synthetic datasets are created to generate pixel-accurate annotations for the invisible parts of objects. Some generate 2D images from synthetic 3D scenes, e.g., DYCE [6] provides natural configuration of objects in indoor scenes and SAIL-VOS[15] provides densely labeled video data extracted from the photo-realistic game GTA-V. These datasets contain natural object boundaries, while being deficient in photo-realistic textures. Others like [19] superimposing objects over other images to create arti-

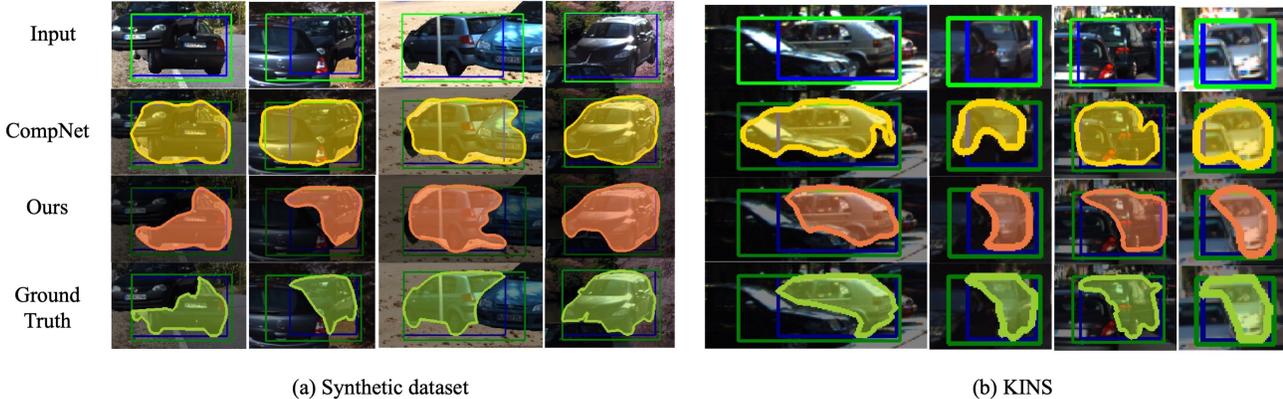


Figure 6: Qualitative results for modal segmentation on KINS and images from our occlusion challenge. The top row show the input images including bounding box annotations. Images in the second row are generated by the baseline CompNet, and the third row shows the results by our CompNet with multi-object ORM. The last row shows the ground truth.

ficial occlusion with real-world textures. For the purpose of studying different types of occlusion challenges, we introduce a dataset with custom artificially generated occlusion scenarios. We crop non-occluded objects from images in KITTI based on their segmentation mask, and place them in images with random backgrounds (Figure 5).

Since only complete visible objects are selected and the exact shape of each object is available, our occlusion challenge provides more accurate annotation for amodal masks compared with the human estimated masks in KINS. Most importantly, the synthetic nature of the dataset allows us to design challenging scenarios occlusion scenarios. In particular, we propose three types of occlusion challenges: 1) The basic and simplest occlusion scenario includes two objects, where one occludes the other. 2) A much more complex occlusion relationship is defined when four objects occlude each other with different amounts partial occlusion. Recovering the occlusion order and modal as well as amodal segmentation requires significant reasoning processes, even for humans. 3) Another challenging scenario is defined when the occluders contain a mixed set of object classes, some of which are known at training time, while some are natural objects that are not part of the training data, such as street signs, bushes, and etc.

## 4.2. Implementation Details

**Baselines.** We implement Mask R-CNN[11] as baseline method for the modal instance segmentation. We further compare to CompositionalNets [28] with and without our proposed multi-object reasoning module. We apply multi-object reasoning either with one reasoning iteration ( $iter=1$ ) or recurrently with two iterations ( $iter=2$ ). Note that the CompositionalNets perform segmentation in a weakly-supervised manner from bounding box annotations only. We compare our method with BBTP [14] and PCNet-M

[35]. BBTP uses a bounding box tightness prior to perform weakly-supervised instance segmentation using box-level annotations. PCNet-M performs amodal mask completion in a self-supervised manner. PCNet-M is trained to recover the amodal mask with a given artificially occluded modal mask. In contrast, our model predicts amodal masks with bounding box supervision only and is capable of handling both known and unknown occluder classes.

**Training setup.** We follow the training strategy as proposed in [16, 28]. CompositionalNets are trained from the feature activations of a ResNeXt-50 [12] model that is pre-trained on ImageNet[4] and fine-tuned on the respective datasets. We set the number of mixture components to  $M = 8$ . We train for 60 epochs using SGD with momentum  $r = 0.9$  and a learning rate of  $lr = 0.01$ .

## 4.3. Instance segmentation under Occlusion

**Modal segmentation.** We report modal instance segmentation performance in the top Tabulars in Table 1 on KINS and Table 2 on our occlusion challenge. Four occlusion levels of objects are defined as: L0: 0%-1%, L1: 1%-30%, L2: 30%-60%, L3:60%-90% of the object area being occluded. To prevent the performance from being affected by a poor bounding box prediction, all models are given the ground truth amodal bounding boxes during training and testing. For the KINS data, we observe that the fully supervised method outperforms weakly-supervised methods. However, our proposed multi-object extension with occlusion reasoning manages to significantly reduce the gap between weakly supervised methods and the fully supervised baseline. We outperform the baseline CompNet performance in every occlusion level, especially in higher occlusion levels by L2(9.6%) and L3 (11.3%) in terms of  $mIoU$ . We observe similar performance patterns on the data for our occlusion challenge. While the CompNet per-

Occ Level	2 Objects					4 Objects					2 Objects + Unknown Occlusion				
	L0	L1	L2	L3	Mean	L0	L1	L2	L3	Mean	L0	L1	L2	L3	Mean
Mask R-CNN	88.2	86.3	69.1	58.2	82.3	88.7	88	74.8	63	78.6	90.5	86.8	72.2	57.1	76.7
CompNet	77.8	67.3	51.0	26.3	66.9	<b>76.7</b>	67.1	50.2	26.1	56.0	<b>78.9</b>	72.2	57.8	36.0	63.6
Ours (iter=1)	<b>78.0</b>	<b>75.3</b>	65.4	45.6	72.9	75.2	<b>72.9</b>	61.9	43.0	65.0	77.9	<b>73.3</b>	<b>62.0</b>	<b>41.7</b>	<b>65.8</b>
Ours (iter=2)	<b>78.0</b>	<b>75.3</b>	<b>65.7</b>	<b>47.2</b>	<b>73.1</b>	75.2	<b>72.9</b>	<b>62.2</b>	<b>44.0</b>	<b>65.3</b>	78.0	<b>73.3</b>	<b>62.0</b>	<b>41.7</b>	<b>65.8</b>

Occ Level	2 Objects					4 Objects					2 Objects + Unknown Occlusion				
	L0	L1	L2	L3	Mean	L0	L1	L2	L3	Mean	L0	L1	L2	L3	Mean
PCNet-M	82.4	81	69.3	47	70	87.2	79.3	63.7	41.3	67.9	-	-	-	-	-
BBTP	<b>80.5</b>	73.6	69.5	72.8	74.1	<b>80.5</b>	71.9	64	66	70.6	<b>83.7</b>	77.3	67.9	60.6	72.4
CompNet	78.0	76.6	75.0	72.1	76.7	77.3	75.4	74.1	71.4	74.8	78.4	<b>78.1</b>	76.1	71.9	76.5
Ours (iter=1)	79.9	<b>80.0</b>	79.2	77.7	<b>79.7</b>	78.6	78.9	78.1	76.6	78.2	78.6	78.0	<b>76.2</b>	<b>72.1</b>	<b>76.6</b>
Ours (iter=2)	79.9	<b>80.0</b>	<b>79.3</b>	<b>78.1</b>	<b>79.7</b>	80.0	<b>80.0</b>	<b>79.3</b>	<b>78.1</b>	<b>79.5</b>	78.5	<b>78.1</b>	<b>76.2</b>	<b>72.1</b>	<b>76.6</b>

Table 2: Modal and amodal instance segmentation on our occlusion challenge (top and bottom respectively). We compare to fully-supervised Mask R-CNN, PCNet-M, and weakly-supervised BBTP, and CompNets with and without ORM. Occlusion levels L0-L3 are defined as: L0: 0%-1%, L1: 1%-30%, L2: 30%-60%, L3: 60%-90% of the object are are occluded. Comparison between different times of occlusion reasoning iteration is also reported. Note that PCNet-M by design cannot handle unknown occlusion, and therefore cannot be applied in the last challenge.

	2 objects		4 objects		2 + unknown	
	Modal	Amodal	Modal	Amodal	Modal	Amodal
NOD	70.5	77.8	58.5	75.2	65.0	76.5
OD	<b>73.1</b>	<b>79.7</b>	<b>65.3</b>	<b>79.5</b>	<b>65.8</b>	<b>76.6</b>

Table 3: Ablation study for order recovery. We compare the modal and amodal instance segmentation results for each occlusion challenge with and without order recovery.

forms similarly for the first and third occlusion challenge, its performance drops significantly when four objects mutually occlude each other compared to the other two scenarios. Our multi-object occlusion reasoning module enables CompNets to close this performance gap. Overall, the multi-object reasoning improves the segmentation results in all occlusion levels and for all challenge scenarios, and in particular for high occlusion levels L2 and L3.

**Amodal segmentation.** We report amodal instance segmentation in the bottom Tabulars in Table 1 and Table 2. Note that the self-supervised PCNet-M requires the modal mask as supervision to learn amodal mask completion. From the results, we observe that our model outperforms all other weakly-supervised methods in all levels of occlusion on the KINS data as well as in the occlusion challenge. We even surpass the mask-supervised PCNet-M in overall performance by 6.5% in *mIoU*.

In summary, with the ability of reasoning about multi-object occlusion, our proposed ORM significantly improves the robustness to occlusion compared with primary CompositionalNet. It achieves accurate instance segmentation in challenging occlusion scenarios (Figure 6. Our weakly-

supervised model even outperforms mask-supervised methods in terms of amodal instance segmentation.

#### 4.4. Ablation study

In Table 3, we verify the effectiveness of the order recovery by evaluating modal and amodal segmentation results on our occlusion challenge. We perform experiments without pair-wise order (NOD), and with our predicted pair-wise order (OD). The results demonstrate the benefit of the order recovery, since per pixel competition cannot always correctly indicate the occluder and the occludee.

## 5. Conclusion

In this paper, we introduced a deep network for multi-object instance segmentation that is robust to occlusion and can be trained from bounding box supervision only. In particular, our network defines a generative model of multiple objects and achieves enhanced robustness through reasoning about multi-object occlusion. We further extended our architecture with an occlusion reasoning module that enables efficient inference in generative models with multiple objects. In particular, it detects erroneous feed-forward predictions and corrects them through reasoning about the occlusion order of objects. Our experiments demonstrate the robustness of our proposed deep network for instance segmentation under occlusion on the KITTI INstance dataset and a dataset with synthetic occluders.

**Acknowledgements.** We gratefully acknowledge funding support from ONR N00014-18-1-2119, ONR N00014-20-1-2206 and the Swiss National Science Foundation (P2BSP2.181713).

## References

- [1] Guang Chen, Fa Wang, Sanqing Qu, Kai Chen, Junwei Yu, Xiangyong Liu, Lu Xiong, and Alois Knoll. Pseudo-image and sparse points: Vehicle detection with 2d lidar revisited by deep learning based methods. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–13, 2020. [1](#)
- [2] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. [2](#)
- [3] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. TensorMask: A foundation for dense object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2061–2069, 2019. [2](#)
- [4] 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. [7](#)
- [5] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. [1](#)
- [6] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. [6](#)
- [7] 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. [3](#)
- [8] Tianshi Gao, Benjamin Packer, and Daphne Koller. A segmentation-aware object detection model with occlusion handling. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1361–1368, 2011. [2](#)
- [9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. [6](#)
- [10] Dileep George, Wolfgang Lehrach, Ken Kansky, Miguel Lázaro-Gredilla, Christopher Laan, Bhaskara Marthi, Xinghua Lou, Zhaoshi Meng, Yi Liu, Huayan Wang, et al. A generative vision model that trains with high data efficiency and breaks text-based captchas. *Science*, 358(6368), 2017. [2](#)
- [11] 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. [2](#), [7](#)
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [7](#)
- [13] Edward Hsiao and Martial Hebert. Occlusion reasoning for object detection under arbitrary viewpoint. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(9):1803–1815, 2014. [2](#)
- [14] 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. *Advances in Neural Information Processing Systems*, 32, 2019. [3](#), [7](#)
- [15] Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, and Alexander G. Schwing. Sail-vos: Semantic amodal instance level video object segmentation - a synthetic dataset and baselines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. [6](#)
- [16] 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. [1](#), [3](#), [4](#), [7](#)
- [17] 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*, pages 1–25, 2020. [1](#), [2](#), [3](#)
- [18] 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. [1](#)
- [19] Ke Li and Jitendra Malik. Amodal instance segmentation. In *European Conference on Computer Vision*, pages 677–693. Springer, 2016. [3](#), [6](#)
- [20] Qizhu Li, Anurag Arnab, and Philip HS Torr. Weakly- and semi-supervised panoptic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 102–118, 2018. [2](#)
- [21] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. [2](#)
- [22] Rui Lu, Feng Xue, Menghan Zhou, Anlong Ming, and Yu Zhou. Occlusion-shared and feature-separated network for occlusion relationship reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10343–10352, 2019. [2](#)
- [23] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele. Exploiting saliency for object segmentation from image level labels. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5038–5047. IEEE, 2017. [2](#)
- [24] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2019. [3](#), [6](#)
- [25] Martin Rajchl, Matthew CH Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Mellisa Damodaram, Mary A Rutherford, Joseph V Hajnal, Bernhard Kainz, et al. Deepcut: Object segmentation from bounding box annotations using convolutional neural net-

- works. *IEEE transactions on medical imaging*, 36(2):674–683, 2016. 2
- [26] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. ” grabcut” interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004. 2
- [27] Fatemehsadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, Stephen Gould, and Jose M Alvarez. Built-in foreground/background prior for weakly-supervised semantic segmentation. In *European Conference on Computer Vision*, pages 413–432. Springer, 2016. 2
- [28] Yihong Sun, Adam Kortylewski, and Alan Yuille. Weakly-supervised amodal instance segmentation with compositional priors. *arXiv preprint arXiv:2010.13175*, 2020. 2, 3, 4, 7
- [29] Joseph Tighe, Marc Niethammer, and Svetlana Lazebnik. Scene parsing with object instances and occlusion ordering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3748–3755, 2014. 2
- [30] Angtian Wang, Yihong Sun, Adam Kortylewski, and Alan L Yuille. Robust object detection under occlusion with context-aware compositionalsnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12645–12654, 2020. 1, 2, 4
- [31] John Winn and Jamie Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, volume 1, pages 37–44. IEEE, 2006. 2
- [32] Mingqing Xiao, Adam Kortylewski, Ruihai Wu, Siyuan Qiao, Wei Shen, and Alan Yuille. Tdmpnet: Prototype network with recurrent top-down modulation for robust object classification under partial occlusion. In *European Conference on Computer Vision*, pages 447–463. Springer, 2020. 2
- [33] Menglong Yang, Yiguang Liu, Longyin Wen, Zhisheng You, and Stan Z. Li. A probabilistic framework for multitarget tracking with mutual occlusions. pages 1298–1305, 2014. 2
- [34] Sangdoon Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *arXiv preprint arXiv:1905.04899*, 2019. 1
- [35] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion, 2020. 2, 7
- [36] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 1
- [37] 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
- [38] 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. 1
- [39] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1464–1472, 2017. 3