WEAKLY-SUPERVISED AMODAL INSTANCE SEGMENTATION WITH COMPOSITIONAL PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Amodal segmentation in biological vision refers to the perception of the entire object when only a fraction is visible. This ability of seeing through occluders and reasoning about occlusion is innate to biological vision but not adequately modeled in current machine vision approaches. A key challenge is that ground-truth supervisions of amodal object segmentation are inherently difficult to obtain. In this paper, we present a neural network architecture that is capable of amodal perception, when weakly supervised with standard (inmodal) bounding box annotations. Our model extends compositional convolutional neural networks (CompositionalNets), which have been shown to be robust to partial occlusion by explicitly representing objects as composition of parts. In particular, we extend CompositionalNets by: 1) Expanding the innate part-voting mechanism in the CompositionalNets to perform instance segmentation; 2) and by exploiting the internal representations of CompositionalNets to enable amodal completion for both bounding box and segmentation mask. Our extensive experiments show that our proposed model can segment amodal masks robustly, with much improved mask prediction qualities compared to state-of-the-art amodal segmentation approaches.

1 INTRODUCTION

In our everyday life, we often observe partially occluded objects. Despite the occluders having highly variable forms and appearances, our human vision system can localize and segment the visible parts of the object, and use them as cues to approximately perceive complete structure of the object. This perception of the object's complete structure under occlusion is referred to as *amodal perception* (Nanay, 2018). Likewise, the perception of visible parts is known to as *modal perception*.

In computer vision, amodal instance segmentation is important to study, both for its theoretical values and real-world applications. Its theoretical similarity to human vision allows for additional insights to the structures of the visual pathway. Also, its real-world importance can be found in the benefits of seeing through the occluder and perceiving partially occluded vehicles in their completeness during autonomous driving. In order to perform amodal segmentation, a vision model must be robust to partial occlusion. Recent works have shown that current deep learning approaches are far less robust than humans at classifying partially occluded objects (Zhu et al., 2019; Kortylewski et al., 2019). In contrast to deep convolutional neural networks (DCNNs), compositional models are much more robust to partial occlusions, as they gained their robustness by mimicking the compositionality of human cognition and sharing similar characteristics with biological vision systems, such as bottom-up object part encoding and top-down attention modulations in the ventral stream (Sasikumar et al., 2018; Roe et al., 2012; Carlson et al., 2011).

Recently, Compositional Convolutional Neural Networks (CompositionalNets) have been proposed as compositional models built upon neural feature activations, which can robustly classify and detect objects under partial occlusions (Kortylewski et al., 2020b). More specifically, Wang et al. (2020) proposed Context-Aware CompositionalNets, which decompose the image into a mixture of object and context.Although Context-Aware CompositionalNets are shown to be robust at detecting objects under partial occlusion, for obvious reasons, they are not sufficient to perform weakly-supervised amodal segmentation. 1) Context-Aware CompositionalNets lack internal priors of the object shape, and therefore cannot perform amodal segmentation. 2) Context-Aware CompositionalNets gather votes from the object parts to vote for an object level classification. This is not sufficient, however, since amodal segmentation requires pixel-level classification. 3) Context-Aware CompositionalNets are high precision models that require the object center to be aligned to the center of the image. However, in practice it is difficult to locate the object center, because only partial bounding box proposals are available for partially occluded objects.

In this work, we propose to build on and significantly extend Context-Aware CompositionalNets in order to enable them to perform amodal instance segmentation robustly with modal bounding box supervision. In particular, we introduce a two-stage model. First, we classify a proposed region and estimate its amodal bounding box via localization of the proposed region on the complete structural representations of the predicted object. Then, we perform per-pixel classification in the estimated amodal region, identifying both visible and invisible regions of the object in order to compute the amodal segmentation mask. Our extensive experiments show that our proposed model can segment amodal masks robustly, with much improved mask prediction qualities compared to current methods under various supervisions. In summary, we make several important contributions in this work:

1. Introduced spatial priors that explicitly encode the prior knowledge of the object's pose and shape in the compositional representation, thus enabling weakly-supervised segmentation.
2. Implemented Partial Classification which maintain the model's accuracy with incomplete object proposals by sampling over all possible spatial placement of the proposal within the internal representation.
3. Implemented Amodal Completion from partial bounding boxes by enforcing symmetry upon the maximum deviation from objective center caused by the spatial placement.
4. Implemented Amodal Segmentation by explicitly classifying the visible and invisible regions within the estimated amodal proposal.

2 Related Work

Robustness to Occlusion In image classification, typical DCNN approaches are significantly less robust to partial occlusions than human vision (Zhu et al., 2019; Kortylewski et al., 2019). Although some efforts in data augmentation with partial occlusion or top-down cues are shown to be effective in reinforcing robustness (DeVries & Taylor, 2017; Xiao et al., 2019), Wang et al. (2020) demonstrate that these efforts are still limited. In object detection, a number of deep learning approaches have been proposed by Zhang et al. (2018) and Narasimhan (2019) for detecting occluded objects; however, these require detailed part-level annotations occlusion reconstruction. In contrast, CompositionalNets, which integrate compositional models with DCNN architecture, are significantly more robust to partial occlusion in image classification under occlusion. Additionally, Context-Aware CompositionalNets, which disentangle its foreground and context representation, are shown to be more robust in object detection under occlusion.

Weakly-supervised Instance Segmentation. Observed in biological vision, pixel-level annotations are not necessary to accomplish object segmentation, since distinguishing between foreground and context in a given region is mainly automatic. Similarly, the feasibility of weakly-supervised instance segmentation in computer vision has been explored. Hsu et al. (2019) achieves figure/ground separation by exploiting the bounding box tightness prior to generate positive and negative bangs based on the sweeping lines of each bounding box. Additionally, Zhou et al. (2018) propose to use image-level annotations to supervise instance segmentation by exploiting class peak responses to enable a classification network for instance mask extraction.

Amodal Perception. One of the first works in amodal instance segmentation was proposed by Li & Malik (2016), with an artificially generated occlusion dataset. Recently, with the release of datasets that contain pixel-level amodal mask annotations, such as KINS and Amodal COCO, further progress has been made (Qi et al., 2019; Zhu et al., 2017). For instance, Zhan et al. (2020) propose a self-supervised network that performs scene de-occlusion, which recovers hidden scene structures without ordering and amodal annotations as supervisions. However, their approach assumes mutual occlusions, thus unfit to perform amodal segmentation when the occluding object is not annotated in the dataset.

3 Weakly Supervised Amodal Segmentation

In Section 3.1, we discuss prior work on CompositionalNets and Context-Aware CompositionalNets. We discuss our extensions to the probabilistic model of Context-Aware CompositionalNets and how they enable weakly-supervised amodal instance segmentation in Section 3.2. Lastly, we discuss the end-to-end training of our model for weakly supervised amodal segmentation in Section 3.3.

Notation. The output of the layer l in the DCNN is referred to as *feature map* $\mathbf{F}^l = \psi(I, \Omega) \in \mathbb{R}^{H \times W \times D}$, where I and Ω are the input image and the parameters of the feature extractor, respectively. *Feature vectors* are vectors in the feature map, $\boldsymbol{f}_{\boldsymbol{i}}^l \in \mathbb{R}^D$ at position $\boldsymbol{i}$, where $\boldsymbol{i}$ is defined on the 2D lattice of $\mathbf{F}^l$ 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: Context-Aware CompositionalNets

CompositionalNets. CompositionalNets, as proposed by Kortylewski et al. (2020a), are DCNN classifiers that are inherently robust to partial occlusion. Their architecture resembles that of a regular DCNN architecture, but the fully connected head is replaced with a differentiable compositional model built upon the feature activations $\mathbf{F}$. They define a probabilistic generative model $p(\mathbf{F}|y)$ with y being the category of the object. Specifically, the compositional model is defined as a mixture of von-Mises-Fisher (vMF) distributions:

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

$$p(\mathbf{F}|\theta_y^m) = \prod_{\boldsymbol{i}} p(\boldsymbol{f_i}|\mathcal{A}_{\boldsymbol{i},y}^m, \Lambda), \;\; p(\boldsymbol{f_i}|\mathcal{A}_{\boldsymbol{i},y}^m, \Lambda) = \sum_k \alpha_{\boldsymbol{i},k,y}^m p(\boldsymbol{f_i}|\lambda_k), \tag{2}$$

Here M is the number of mixtures of compositional models per each object category and ν_m is a binary assignment variable that indicates which mixture component is active. $\Theta_y = \{\theta_y^m = \{\mathcal{A}_y^m, \Lambda\}|m = 1, \ldots, M\}$ are the overall compositional model parameters for the category y and $\mathcal{A}_y^m = \{\mathcal{A}_{\boldsymbol{i},y}^m|\boldsymbol{i} \in [H, W]\}$ are the parameters of the mixture components at every position $\boldsymbol{i}$ on the 2D lattice of the feature map $\mathbf{F}$. In particular, $\mathcal{A}_{\boldsymbol{i},y}^m = \{\alpha_{\boldsymbol{i},k,y}^m|k = 1, \ldots, K\}$ are the vMF mixture coefficients and $\Lambda = \{\lambda_k = \{\sigma_k, \mu_k\}|k = 1, \ldots, K\}$ are the parameters of the vMF mixture distributions. Note that K is the number of parameters in the vMF mixture distributions and the sum across all K vMF mixture coefficients, $\sum_{k=0}^{K} \alpha_{i,k,y}^m = 1$.

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

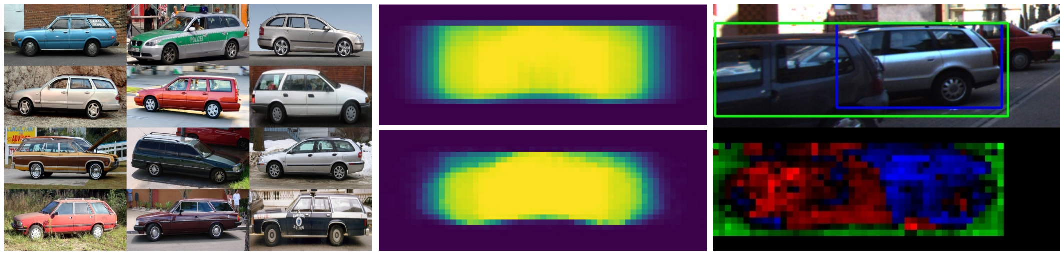
where $Z(\sigma_k)$ is the normalization constant. The model parameters $\{\Omega, \{\Theta_y\}\}$ can be trained end-to-end as described in Kortylewski et al. (2020a).

Context awareness. As introduced by Wang et al. (2020), context-aware CompositionalNets expand on the standard CompositionalNets and explicitly separates the representation of the context from the object by representing the feature map $\mathbf{F}$ as a mixture of two.

$$p(\boldsymbol{f_i}|\mathcal{A}_{\boldsymbol{i},y}^m, \chi_{\boldsymbol{i},y}^m, \Lambda) = \omega \; p(\boldsymbol{f_i}|\chi_{\boldsymbol{i},y}^m, \Lambda) + (1 - \omega) \; p(\boldsymbol{f_i}|\mathcal{A}_{\boldsymbol{i},y}^m, \Lambda). \tag{4}$$

Here, the object representation is disentangled into the foreground representation $\mathcal{A}_{\boldsymbol{i},y}^m$ and context representation $\chi_{\boldsymbol{i},y}^m$. The scalar ω is a prior that controls the trade-off between context and object, which is fixed a priori at test time. It is shown that although context is helpful in detecting objects under partial occlusions, relying too strongly on context can be misleading when objects are strongly occluded, leading to a relatively high object confidence in background regions.

In order to achieve foreground/context disentanglement, training images are segmented into either object or context based on the contextual feature centers, $\boldsymbol{e}_q \in \mathbb{R}^D$, learned through available bounding box annotation. Here, the assumption is that any feature with receptive field outside of the bounding boxes is considered to be contextual features. Thus, a dictionary of context feature centers $E = \{\boldsymbol{e}_q \in \mathbb{R}^D|q = 1, \ldots, Q\}$ can be learned through clustering the population of randomly



(a) Object Cluster (b) Compositional Prior (c) Amodal Completion

Figure 1: Amodal Completion with Compositional Priors. (a) demonstrates the object cluster that the compositional prior in (b) is trained on. We observe an optimization of the compositional prior from iteration 0 (top) to iteration 2 (bottom). Lastly, (c) shows amodal estimation from modal (blue) to amodal (green) bounding box. The pixel-level labels, $\mathcal{F}$, $\mathcal{O}$, and $\mathcal{C}$, are depicted as blue, red, green, scaled by their responses, respectively.

extract contextual features using K-means++ (Arthur & Vassilvitskii, 2007). Finally, the binary classification of the feature vector $\boldsymbol{f_i}$ to either foreground, $\mathcal{F}$, or context, $\mathcal{C}$, is determined such that:

$$\boldsymbol{f_i} = \begin{cases} \mathcal{F}, & \text{if } \max_q[(\boldsymbol{e}_q^T \boldsymbol{f_i})/(||\boldsymbol{e}_q||||\boldsymbol{f_i}||)] < 0.5 \\ \mathcal{C}, & \text{otherwise} \end{cases} \tag{5}$$

3.2 Weakly-supervised Amodal Instance Segmentation

Segmentation with Spatial Compositional Priors. The Context-Aware CompositionalNets, as proposed by Wang et al. (2020), generates *object-level* predictions, i.e. class labels, by gathering votes from local part detectors. Our objective, on the other hand, is to generate *pixel-level* predictions to perform instance segmentation. A simple strategy would be to use the ratio between the context and the foreground likelihood from Equation 4. While this can give reasonable results shown by Kortylewski et al. (2020c), a major limitation of this approach is that the prior ω is independent of the position p and the object pose m. However, the likelihood of a feature being part of the context is clearly dependent on the shape of the object and hence depends on these variables.

Therefore, we propose a spatial prior $p(i|m, y)$ to explicitly encode the prior knowledge of the object pose m and shape in the representation model. Seen in Figure 1, the compositional prior $p(i|m, y)$ is defined over all position $\boldsymbol{i}$ and for every mixture m. Note how the prior clearly resembles the object shape and 3D pose. Formally, we can learn $p(i|m, y)$ by computing the average foreground segmentation of each training image that is used to train the mixture model m of class y. We extend the probabilistic compositional model to incorporate the learned spatial priors as a mixture model:

$$p(\boldsymbol{f}_i|\mathcal{A}^m_{i,y}, \chi^m_{i,y}, \Lambda) = (1 - p(i|m, y))\; p(\boldsymbol{f_i}|\chi^m_{\boldsymbol{i},y}, \Lambda) + p(i|m, y)\; p(\boldsymbol{f_i}|\mathcal{A}^m_{\boldsymbol{i},y}, \Lambda). \tag{6}$$

To segment $\mathbf{F}$ into foreground $\mathcal{F}$ and context $\mathcal{C}$, we use the ratio between the two components:

$$\boldsymbol{f_i} = \begin{cases} \mathcal{F}, & \text{if } \log\left[p(i|m, y)\; p(\boldsymbol{f_i}|\mathcal{A}^m_{\boldsymbol{i},y}, \Lambda)\right] - \log\left[(1 - p(i|m, y))\; p(\boldsymbol{f_i}|\chi^m_{\boldsymbol{i},y}, \Lambda)\right] > 0 \\ \mathcal{C}, & \text{otherwise} \end{cases} \tag{7}$$

The spatial prior also allows us to estimate the context separation during training more accurately in an EM-type manner. In particular, we perform an initial segmentation following the approach proposed by Wang et al. (2020). Subsequently, we learn the spatial prior and update the initial segmentation using Equation 7. As illustrated in the Figure 1b, the spatial prior is optimized in both its tightness and confidence through the iterative updates, since utilizing explicit prior knowledge of the object shape outperforms the contextual features at instance segmentation.

Maximum likelihood Alignment of Partial Feature Maps. As pointed out by Wang et al. (2020), CompositionalNets are high-precision models because they assume that the object is aligned to the center of the compositional model. However, this assumption is only valid if the amodal bounding box is available and hence would not work when a bounding box proposal only contains a part of the object. This poses as a substantial barrier to apply it to amodal perception, since targeted objects are occluded and amodal bounding boxes may not be avaliable during training or inference. Therefore, we propose to obtain the maximum likelihood alignment of feature maps by searching over the spatial placement of F on the compositional representation θ_y^m. This will remove the alignment constraint and, consequently, allow us to leverage partial proposals for amodal perception.

$$p(\mathsf{F}|\Theta_y) = \sum_m \nu_m \max_{\boldsymbol{i}} p(\mathsf{F}^{\boldsymbol{d}}|\theta_y^m), \quad \boldsymbol{d} \in [0, H' - H + 1] \times [0, W' - W + 1] \tag{8}$$

Here, $\mathsf{F}^{\boldsymbol{d}}$ denotes F with a particular zero padding that aligns the top left corner of F to the position $\boldsymbol{d}$ defined on the 2D lattice of the internal compositional representation θ_y^m, where (H, W) and (H', W') being the spatial dimension of F and θ_y^m, respectively.

Shown the Figure 1c, by maximizing the likelihood of F on the representation, we would be able to localize correctly to the compositional representation. As we will show in the next section, such localization $\boldsymbol{d}$ is used to estimate the amodal region, combined with the compositional priors.

Amodal Bounding Box Completion. After obtaining the corresponding coordinate $\boldsymbol{d}$ and representation model $\mathcal{A}_{\boldsymbol{i},y}^m$, we proceed to estimate the complete structure of the object and perform amodal completion on the bounding box level. The estimation of amodal bounding box depends both on the compositional prior and the localization of F on the representation. For the rest of this paragraph, we shift the global axis from the image to the representation. The object center, in this case, is trivially defined as the center of the representation, $\boldsymbol{c} = [\frac{H'}{2}, \frac{W'}{2}]$. Assuming that any bounding box is defined in a form $[\boldsymbol{a}, \boldsymbol{b}]$ where $\boldsymbol{a}$ and $\boldsymbol{b}$ are the top left and bottom right of the box, respectively. We proposed the estimation of amodal bounding box $\mathcal{B}$ from modal bounding box $B = [\boldsymbol{d}, \boldsymbol{d} + [H, W]]$:

$$\mathcal{B} = [\boldsymbol{c} - \boldsymbol{k}, \boldsymbol{c} + \boldsymbol{k}] \text{ ,where,} \tag{9}$$

$$\boldsymbol{k} = \begin{cases} \boldsymbol{c} - \boldsymbol{d}, & \text{if } ||\boldsymbol{c} - \boldsymbol{d}|| > ||\boldsymbol{d} + [H, W] - \boldsymbol{c}|| \\ \boldsymbol{d} + [H, W] - \boldsymbol{c}, & \text{otherwise} \end{cases} \tag{10}$$

Here, $\boldsymbol{k}$ denotes the maximum displacement vector observed at localization $\boldsymbol{d}$. By applying $\boldsymbol{k}$ symmetrically to the object center $\boldsymbol{c}$, an amodal estimation of the object region $\mathcal{B}$ is generated.

Amodal Instance Segmentation with CompositionalNet. As we discussed above, segmentation with CompositionalNets is treated as per-pixel binary classification between foreground $\mathcal{F}$ and context $\mathcal{C}$ on the feature layer F. In order to perform amodal instance segmentation, both the visible and invisible mask of the object must be explicitly obtained. Therefore, we propose a third category for the per-pixel classification, $\mathcal{O}$, denoting the occluded pixels of the object.

Reasonably, these occluded pixels of the object have high compositional prior and low likelihood probability. Since we view occluded regions as unexplainable to our compositional representation instead of explicit occluders, we propose an outlier model, $\boldsymbol{o} \in \mathbb{R}^K$, such that its representation is broadly defined over the entire dataset, in an attempt to model any features vector unexplainable to the compositional representation. Here, $\boldsymbol{o}$ has the same dimensions as a compositional representation at a particular position $\boldsymbol{i}$, namely $\mathcal{A}_{\boldsymbol{i},y}^m$. Thus, $p(\boldsymbol{f_i}|\boldsymbol{o}, \Lambda)$ is calculated the same way as $p(\boldsymbol{f_i}|\mathcal{A}_{\boldsymbol{i},y}^m, \Lambda)$. This way, occlusion can be properly modeled by a high activation of the outlier model, compared to the compositional and context models. By combining the high compositional prior and low likelihood probability together, we formulate the probability that any feature vector $\boldsymbol{f_i}$ is classified as an occluded object $\mathcal{O}$ as below:

$$p(\boldsymbol{f_i} = \mathcal{O}) = p(i|m, y)\ p(\boldsymbol{f_i}|\boldsymbol{o}, \Lambda) \tag{11}$$

Since amodal segmentation is defined by the union of visible and invisible masks, amodal segmentation can be modeled as $\{p \mid \boldsymbol{f_i} = \mathcal{F}\} \cup \{p \mid \boldsymbol{f_i} = \mathcal{O}\}$.

$$\boldsymbol{f_i} = \begin{cases} \mathcal{O}, & \text{if } \log p(\boldsymbol{f_i} = \mathcal{O}) - \log \Big[\max p(\boldsymbol{f_i} = \mathcal{F}), p(\boldsymbol{f_i} = \mathcal{C}) \Big] > 0 \\ \mathcal{F}, & \text{if } \log p(\boldsymbol{f_i} = \mathcal{F}) - \log \Big[\max p(\boldsymbol{f_i} = \mathcal{O}), p(\boldsymbol{f_i} = \mathcal{C}) \Big] > 0 \\ \mathcal{C}, & \text{otherwise} \end{cases} \tag{12}$$

3.3 End-to-End Training

Overall, the trainable parameters of our models are $T = \{\mathcal{A}_y\ \chi_y\}$, with ground truth modal bounding box B and label y as supervision. The loss function has two main objectives: 1) improve classification accuracy under occlusion ($\mathcal{L}_{cls}$). 2) promote maximum likelihood for compositional and context representations ($\mathcal{L}_{rep}$). 2) improve amodal segmentation quality ($\mathcal{L}_{seg}$).

Training Classification with Regularization. We optimize the parameters jointly using SGD, where $\mathcal{L}_{class}(\hat{y}, y)$ is the cross-entropy loss between the model output $\hat{y}$ and the true class label y.

Training the Generative Model with Maximum Likelihood. Here, we use $\mathcal{L}_{rep}$ to enforce a maximum likelihood for both the compositional and context representation over the dataset. Note that $m^{\uparrow}$ denote the mixture assignment that is inferred in the forward process and the outlier model is learned a priori and then fixed.

$$\mathcal{L}_{rep}(F, \mathcal{A}_y, \chi_y) = -\sum_{\boldsymbol{i}} \log\Big[\sum_{k} \theta^{m^{\uparrow}}_{\boldsymbol{i},k,y} p(f_{\boldsymbol{i}}|\lambda_k)\Big] \text{ ,where,} \tag{13}$$

$$\Theta^{m}_{\boldsymbol{i},y} = p(\mathcal{A}^{m}_{\boldsymbol{i},y})\ \mathcal{A}^{m}_{\boldsymbol{i},y} + (1 - p(\mathcal{A}^{m}_{\boldsymbol{i},y}))\ \chi^{m}_{\boldsymbol{i},y} \tag{14}$$

Training Segmentation with Regularization. This loss function that is based on the bounding box tightness prior is proposed by Hsu et al. (2019). Since $\mathcal{L}_{cls}$ by itself would motivate representations to focus on specific regions of the object instead of the complete object, $\mathcal{L}_{seg}$ proves to be significant, as it motivate representations to have a consistent explainability over the entire object.

$$\mathcal{L}_{seg}(\hat{m}, B) = \sum_{b} -\log \max \hat{m}(b) - \sum_{b'} \log(1 - \max \hat{m}(b')) + \delta \sum_{(\boldsymbol{i},\boldsymbol{i}') \in \epsilon} (\hat{m}(\boldsymbol{i}) - \hat{m}(\boldsymbol{i}'))^2 \tag{15}$$

Here, denote $\hat{m}$ as the predicted mask in image space, and B as the bounding box as supervision. b is the set containing sweep rows and columns within the bounding box, while b' is the set containing sweep rows and columns directly outside the bounding box. Additionally, ϵ is the set containing all neighboring pixel pairs, while δ controls the trade-off between the two loss terms. Intuitively, $\mathcal{L}_{seg}$ is composed of two parts. First part is referred as the unary term, as it enforces every row or columns of pixels within the bounding box to contain at least one pixel that is recognized as a part of the predicted mask, while discouraging mask predictions outside of the bounding box. Second part is referred as the pairwise term, as it enforces pair-wise smoothness within the predicted mask.

End-to-end training. We train all parameters of our model end-to-end with the overall loss function:

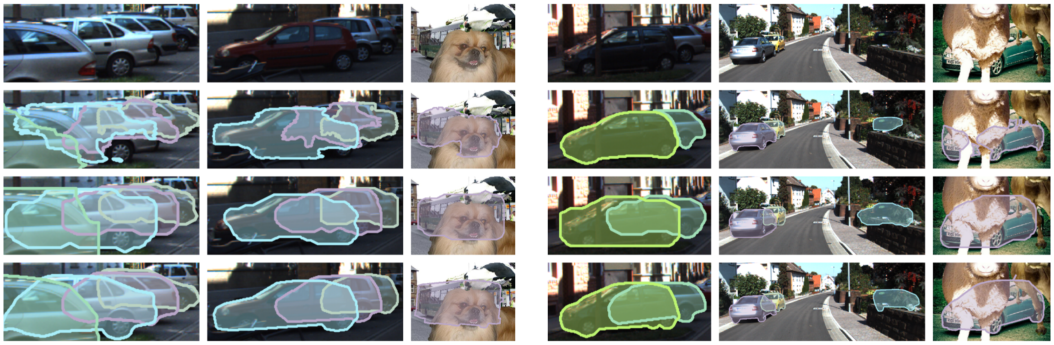
$$\mathcal{L} = \mathcal{L}_{cls}(\hat{y}, y) + \gamma_1 \mathcal{L}_{rep}(F, \mathcal{A}_y, \chi_y) + \gamma_2 \mathcal{L}_{seg}(\hat{m}, B) \tag{16}$$

while γ_1 and γ_2 controls the trade-off between the loss terms.

4 Experiments

We perform experiments on semi-supervised amodal instance segmentation under both artificially-generated and real-world occlusion.

Datasets. While it is important to evaluate the approach on real images of partially occluded objects, simulating occlusion enables us to quantify the effects of partial occlusion more accurately. For the artificial dataset, we evaluated our approach on the *OccludedVehiclesDetection* dataset proposed by Wang et al. (2020). We remove the train category from evaluation due to the inaccurate mask annotations that only pertains to one segment of the train. The occlusion exists in both the object and its context by objects such as humans, animals and plants cropped from the MS-COCO dataset. The loss of contextual information increases the difficulty of amodal segmentation as the overall amodal structure of the object is removed. The *OccludedVehiclesDetection* contains 9 occlusion levels along two dimensions, which include three levels of object occlusion: FG-L1: 20-40%, FG-L2: 40-60% and FG-L3: 60-80% of the object area occluded, and three levels of context occlusion around the object: BG-L1: 0-20%, BG-L2: 20-40% and BG-L3: 40-60% of the contextual area occluded.


(a) BBTP Comparison

(b) PCNet-M Comparison

Figure 2: Qualitative Amodal Instance Segmentation Results. From the top to the bottom row, we present the raw image, BBTP/PCNet-M predictions, our method's predictions, and Ground Truth.

For the realistic dataset, we evaluate our approach on the *KINS* dataset proposed by Qi et al. (2019). Similar to the *OccludedVehiclesDetection* dataset, we split the objects into 3 object occlusion levels: FG-L1: 1-30%, FG-L2: 30-60% and FG-L3: 60-90%. We restrict the scope of the evaluation to vehicles that have a minimum amodal height of 50 pixels, as the significance of the segmentation quality decreases when the resolution of object reduces to too low.

CompositionalNets. We implement the end-to-end training of our proposed model with the following parameter settings: training minimizes loss described in Equation 3.3, with $\gamma_1 = 3$, $\gamma_2 = 1$ and $\delta = 0.05$. We applied the Adam Optimizer proposed by Kingma & Ba (2014) with learning rate $lr_{=}1 \cdot 10^{-3}$. Our proposed model is trained for a total of 1 epoch of 10000 iterations. The training costs in total of 2 hours on a machine with 1 NVIDIA TITAN Xp GPUs.

BBTP, proposed by Hsu et al. (2019), explores the bounding box tightness prior as its mechanism to generate segmentation mask with weak supervision. BBTP is trained for 20000 iterations, with a learning rate/decay, $lr = 1 \cdot 10^{-2}$, $lr_{decay} = 1 \cdot 10^{-4}$. It is trained with non-occluded objects with amodal bounding boxes. Due to its weakly supervised nature, it is not possible to introduce occluder information into training, thus augmented training would not be plausible to implement.

PCNet-M , proposed by Zhan et al. (2020), learns amodal completion from artificially placing other objects in the dataset as occluders on the objects in a self-supervised manner given modal segmentation masks. It is trained for 20000 iterations, with a learning rate, $lr = 1 \cdot 10^{-3}$, $lr_{decay} = 1 \cdot 10^{-4}$. Mask RCNN, proposed by He et al. (2017), serves as a modal segmentation network for PCNet-M. It is trained for 40000 iterations, with a learning rate/decay, $lr = 1 \cdot 10^{-3}$, $lr_{decay} = 5 \cdot 10^{-4}$. Similarly, it is also trained with non-occluded objects. Due to its self-supervised amodal completion, augmented training is implied within the model's construction. Therefore, PCNet-M is viewed as the fully supervised approach as oppose to our weakly supervised model.

Evaluation. As seen in the *KINS* dataset, the occlusion levels of objects are severely disproportional, observing over 62% of the objects are non-occluded and less than 8% of objects are in the highest occlusion level. Therefore, in order to examine the mask prediction quality as a function of occlusion levels, we evaluate with region proposals as supervision, in order to remove the bias to non-occluded objects and separate objects into subsets based on their occlusion level during evaluation. Since BBTP is only trained on complete amodal bounding boxes, it is unreasonable to evaluate it with modal bounding box. Therefore, it will be evaluated with amodal bounding boxes. On the other hand, since PCNet-M focuses its attention on self-supervision without occlusion annotation during training, PCNet-M will be evaluated with modal bounding boxes. In the end, we evaluate our approach in the same setting as both models separately.

4.1 Amodal Segmentation under Simulated Occlusion

Table 1 and Figure 2a shows the results of the tested models on the OccludedVehiclesDetection dataset.

Table 1: Amodal Segmentation is evaluated on the OccludedVehiclesDetection Dataset with meanIoU as the performance metric. For supervision, a and m denotes the amodal and modal bounding box, respectively. Also, * denotes the ground truth occluder segmentation given as supervision.

FG Occ. Level	-	**0**	**1**			**2**			**3**			Mean
BG Occ. Level	-	-	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	-
PCNet-M	m*	**79.6**	**75.5**	**72.8**	**69.9**	**69.4**	**65.2**	**61.1**	**61.7**	56.4	50	**66.2**
ours	m	67.7	67.3	66.2	65.4	64.2	62.6	60.8	59.7	**57.6**	**53.4**	62.5
BBTP	a	68.4	62.2	60.7	60	57.6	54.1	51.6	54.2	48.5	43.8	56.1
ours	a	**68.7**	**67.1**	**66.9**	**66.9**	**66.3**	**66.2**	**66.1**	**65.4**	**65.3**	**65.2**	**66.4**

Table 2: Amodal Segmentation is evaluated on the KINS Dataset with meanIoU as the performance metric. Note that BG Occ. Level is omitted due to physical constraints of realistic occluders.

FG Occ. Level	-	**0**	**1**	**2**	**3**	Mean
PCNet-M	m	**89**	**75.1**	49.1	30.5	60.9
ours	m	72.1	70.4	**68.4**	**52.7**	**65.9**
BBTP	a	73	70.3	66.6	64.7	68.7
ours	a	**77.4**	**74.9**	**78.1**	**76.3**	**76.7**

PCNet-M. First, it is essential to note that PCNet-M requires the ground truth occluder segmentation mask. Furthermore, PCNet-M cannot reason about partial occlusions and amodal completion if the occluder category is unknown during training. In the case of the OccludedVehiclesDetection dataset, the occluders class labels are not given, thus it becomes necessary to given additional information to the PCNet-M. In contrast, our approach does not require any additional information regarding to the occluder during inference. From the results in Table 1we can observe that, although PCNet-M is trained with mask supervision, our approach is able to outperform the PCNet-M in amodal segmentation at higher object occlusions.

BBTP. The proposed model is able to outperform BBTP in amodal segmentation across all occlusion settings, including non-occluded objects. Hence our modal achieves state-of-the-art performance at weakly supervised amodal segmentation.

4.2 Amodal Segmentation under Realistic Occlusion

Table 2 shows the results of the tested models on the KINS dataset and Figure 2 b refers to the qualitative results. Notably, a similar trend observed in the OccludedVehiclesDetection dataset is found in the KINS dataset with realistic occlusion.

PCNet-M. Seen in the table, PCNet-M outperforms CompositionalNets in lower levels of occlusion, but fails to perform amodal completion over large occluded regions in high level occlusion cases.

BBTP. Similarly observed as above, CompositionalNets exceeds in segmentation performance across all occlusion levels compared to BBTP.

5 Conclusion

In this work, we studied the problem of weakly-supervised amodal instance segmentation with partial bounding box annotations only. We made the following contributions to advance the state-of-the-art in weakly-supervised amodal instance segmentation: 1) We extend the Context-Aware **CompositionalNets with innate spatial priors** of the object shape to enable weakly-supervised amodal instance segmentation. 2) We enable CompositionalNets to **predict the amodal bounding box of an object** based on a modal (partial) bounding box, via maximum likelihood alignment of the partial feature representation with the internal object representation.3) We show that **deep networks are capable of amodal perception**, when they are augmented with compositional and spatial priors. Furthermore, we demonstrate that deep networks can learn the necessary knowledge in a weakly supervised manner from bounding box annotations only.

REFERENCES

D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. *In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2007.

E.T. Carlson, R.J. Rasquinha, K. Zhang, and C.E. Connor. A sparse object coding scheme in area v4. *Current Biology*, 2011.

Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

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. dAlché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 6586–6597. Curran Associates, Inc., 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Adam Kortylewski, Qing Liu, Huiyu Wang, Zhishuai Zhang, and Alan Yuille. Combining compositional models and deep networks for robust object classification under occlusion. *arXiv preprint arXiv:1905.11826*, 2019.

Adam Kortylewski, Ju He, Qing Liu, and Alan Yuille. Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion. *IEEE Conference on Computer Vision and Pattern Recognition*, 2020a.

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*, pp. 8940–8949, 2020b.

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. *arXiv preprint arXiv:2006.15538*, 2020c.

Ke Li and Jitendra Malik. Amodal instance segmentation. In *European Conference on Computer Vision*, pp. 677–693. Springer, 2016.

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

N. Dinesh Reddy Minh Vo Srinivasa G. Narasimhan. Occlusion-net: 2d/3d occluded keypoint localization using graph networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

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.

A.W. Roe, L. Chelazzi, C.E. Connor, B.R. Conway, I. Fujita, J.L. Gallant, H. Lu, and W. Vanduffel. Toward a unified theory of visual area v4. *Neuron*, 2012.

D. Sasikumar, E. Emeric, V. Stuphorn, and C.E. Connor. First-pass processing of value cues in the ventral visual pathway. *Current Biology*, 2018.

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*, pp. 12645–12654, 2020.

Mingqing Xiao, Adam Kortylewski, Ruihai Wu, Siyuan Qiao, Wei Shen, and Alan Yuille. Tdapnet: Prototype network with recurrent top-down attention for robust object classification under partial occlusion. *arXiv preprint arXiv:1909.03879*, 2019.

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.

Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Occlusion-aware r-cnn: detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 637–653, 2018.

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*, pp. 3791–3800, 2018.

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.

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.