Sequential Amodal Segmentation via Cumulative Occlusion Learning

Jiayang Ao[1]
https://jiayangao.github.io/
Qiuhong Ke[2]
qiuhong.ke@monash.edu
Krista A. Ehinger[1]
kehinger@unimelb.edu.au

[1] The University of Melbourne
Parkville, VIC, Australia
[2] Monash University
Clayton, VIC, Australia

Abstract

To fully understand the 3D context of a single image, a visual system must be able to segment both the visible and occluded regions of objects, while discerning their occlusion order. Ideally, the system should be able to handle any object and not be restricted to segmenting a limited set of object classes, especially in robotic applications. Addressing this need, we introduce a diffusion model with cumulative occlusion learning designed for sequential amodal segmentation of objects without specifying their categories. This model iteratively refines the prediction using the cumulative mask strategy during diffusion, effectively capturing the uncertainty of invisible regions and adeptly reproducing the complex distribution of shapes and occlusion orders of occluded objects. It is akin to the human capability for amodal perception, i.e., to decipher the spatial ordering among objects and accurately predict complete contours for occluded objects in densely layered visual scenes. Experimental results across three amodal datasets show that our method outperforms established baselines. The code is available at github.com/saraao/SAS.

1 Introduction

Robots often encounter unfamiliar objects in ever-changing unstructured environments such as warehouses or homes [29]. These scenarios require systems capable of manipulating objects based on their complete shape and occlusion relationships rather than their visibility or category [2, 7, 31]. However, most state-of-the-art amodal segmentation methods [1, 8, 14, 30], which are usually constrained by the need for class-specific data, struggle to generalize to unseen objects and are susceptible to misclassification.

We introduce a novel diffusion model for sequential amodal segmentation that does not rely on object categories. Our approach transcends traditional single or dual-layer prediction limitations [11, 16, 21] by enabling the simultaneous segmentation of unlimited object layers in an image. In addition, our framework generates multiple plausible amodal masks for each object from a single input image, contrasting with prior approaches that depend on multiple ground truths to achieve varied results [9, 24, 32]. Tailored to the amodal task, our method requires only a single ground truth per object during training to capture the diversity of

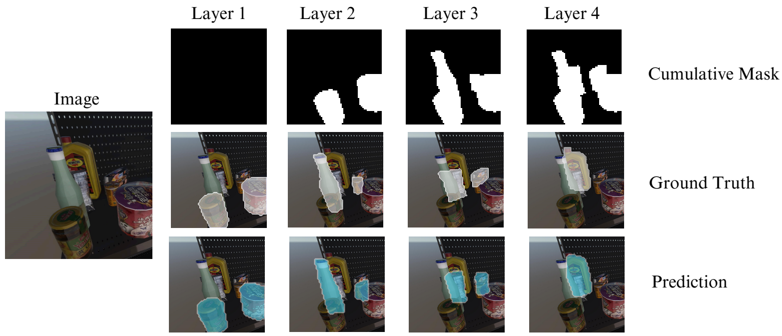

Figure 1: The cumulative mask and amodal mask predictions for an input image. Our method can generate reliable amodal masks layer by layer and allows multiple objects per layer.

occlusions, overcoming the limitations of existing amodal datasets that typically provide only one annotation per object and neglect the variability in invisible regions.

Our framework takes an RGB image as input and sequentially predicts the amodal masks for each object, as illustrated in Fig. 1. The iterative refinement process of our proposed algorithm, inspired by human perception mechanisms for invisible regions [26], leverages preceding identified items to infer subsequent occluded items. Specifically, it employs a cumulative mask, which aggregates the masks of previously identified objects. This strategy allows the model to maintain a clear record of areas already segmented, directing its focus toward unexplored regions. By focusing the prediction effort on uncertain or occluded regions, our approach improves the accuracy and reliability of the amodal segmentation process. We validate our approach through comprehensive ablation studies and performance benchmarking across three amodal datasets, demonstrating its superiority in handling complex sequential amodal segmentation challenges.

The main contributions of our work are:

- A new sequential amodal segmentation method capable of predicting unlimited layers of occlusion, enabling occlusion modelling in complex visual scenes.
- Occluded shape representation which is not based on labelled object categories, enhancing its applicability in diverse and dynamic settings.
- A diffusion-based approach to generating amodal masks that captures the uncertainty over occluded regions, allowing for diverse segmentation outcomes.

2 Related Work

Amodal segmentation with order perception requires segmentation of the entire objects by including both visible and occluded regions while explicitly resolving the layer order of all objects in the image. Establishing layering of objects allows for a comprehensive understanding of the scene and the spatial relationships between objects, which is essential for tasks such as autonomous driving, robot grasping, and image manipulation [2, 13, 38]. Current amodal segmentation methods mainly assess occlusion states of individual objects [6, 21, 25, 28] or between pairs [2, 11, 35], but tend to ignore the global order in a complex scene, such as the relationship between independent groups. While some work [1, 38] has begun to address amodal segmentation with perceptible order, they fall short for class-agnostic applications due to design constraints on category-specific dependencies.

Class-agnostic segmentation detects masks without relying on pre-learned category-specific knowledge. It is vital for scenarios where comprehensive labelling is resource-intensive or when encountering unseen categories [22, 29]. However, amodal segmentation approaches usually depend on predefined class labels and thus have limited ability to handle unknown objects [14, 18]. While there are a few methods which consider the class-agnostic amodal segmentation, [2] is for RGB-D images with depth data rather than RGB images, [5] relies on the bounding box of the object as an additional input to predict amodal masks, [39] treats amodal masks prediction and ordering as separate tasks thus designs the methods individually, and other requires additional inputs for prediction such as visible mask [19, 37].

Segmentation with diffusion models has recently attracted interest as its ability to capture complex and diverse structures in an image that traditional models might miss [4, 15, 33, 34]. Particularly in medical imaging, diffusion models are used to generate multiple segmentation masks to simulate the diversity of annotations from different experts [9, 24, 32, 36]. However, these methods are designed for the visible part of images and do not adequately address the diversity of predictions required for the hidden part of objects.

3 Problem Definition

Our goal is to amodally segment multiple overlapping objects within an image without object class labels, while determining the occlusion order of these objects. Specifically, for a given RGB image I, the goal of our sequential amodal segmentation approach is two-fold. First, to produce a collection of amodal segmentation masks $\{M_i\}_{i=1}^N$, where each mask M_i represents the full extent of the corresponding object O_i within the scene—this includes both visible and occluded regions. Second, to assign a layer ordering $\{L_i\}_{i=1}^N$ to these objects based on their mutual occlusions, thereby constructing an occlusion hierarchy.

The layer variable L_i adheres to the occlusion hierarchy defined by [1]. The bi-directional occlusion relationship $Z(i,j)$ indicates if O_i is occluded by O_j, given by:

$$Z(i,j) = \begin{cases} 1, & \text{if object } O_i \text{ is occluded by object } O_j, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The set S_i comprises indices of those objects occluding O_i, is defined by $S_i = \{j | Z(i,j) = 1\}$. Subsequently, the layer ordering L_i for each object O_i is computed based on:

$$L_i = \begin{cases} 1, & \text{if } S_i = \emptyset, \\ 1 + \max_{j \in S_i} L_j, & \text{otherwise.} \end{cases} \tag{2}$$

The ultimate goal is to derive an ordered sequence of amodal masks $\tau = \langle M_1, \ldots, M_N \rangle$ that correctly represents the object layers in image I.

4 Methodology

The architecture of our proposed model is shown in Fig. 2. Details on the architectural components, the cumulative guided diffusion model and the cumulative occlusion learning

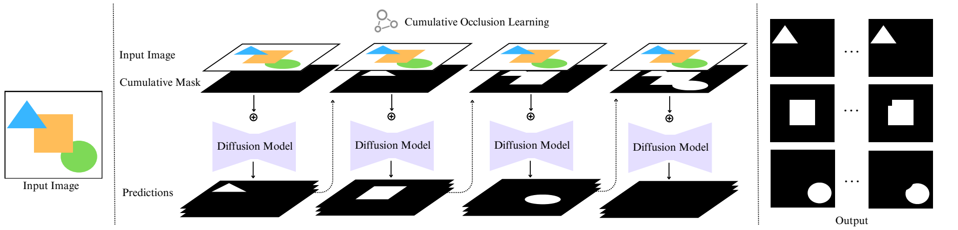

Figure 2: Architecture of our model. Our model receives an RGB image as input and predicts multiple plausible amodal masks layer-by-layer, starting with the unoccluded objects and proceeding to deeper occlusion layers. Each layer's mask synthesis receives as input the cumulative occlusion mask from previous layers, thus providing a spatial context for the diffusion process and helping the model better segment the remaining occluded objects.

algorithm are discussed in Sections 4.1 and 4.2, respectively.

4.1 Diffusion-based Framework

Cumulative mask. We introduce the cumulative mask—a critical innovation that incorporates the spatial structures of objects, facilitating the understanding of both visible and occluded object parts. The cumulative mask aggregates the masks of all objects which are in front of (and potentially occluding) the current layer. Specifically, the cumulative mask for an object O_i with layer order L_i encompasses the masks of all objects with a layer order lower than L_i, thereby representing the cumulative occlusion up to that layer. For each object O_i with its amodal mask M_i and layer order L_i, the cumulative mask CM_i is formalized as:

$$\mathrm{CM}_i = \bigcup_{\{j|L_j<L_i\}} M_j, \tag{3}$$

where $\bigcup$ denotes the union operation, CM_i is the cumulative mask for object O_i, M_j are the masks of objects with a lower layer order L_j than that of O_i, reflecting the cumulative occlusion encountered up to object O_i. $\mathrm{CM} = \varnothing$ denotes no prior occlusion and is used for the fully visible objects in L_1.

Cumulative guided diffusion. We enhance denoising diffusion probabilistic models (DDPMs) [10, 20] to address the unique challenge of understanding occluded regions for amodal segmentation. The diffusion process is informed by a static representation of the input image and the cumulative mask from previous layers. The diffusion process generates an amodal mask for the current layer's objects, which is then added to the cumulative occlusion mask to generate the next layer.

Following the standard DDPMs implementation [10], the diffusion process is modelled as a Markov chain. The forward process q at time t evolves from the previous step $t-1$ is:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}), \tag{4}$$

where x_t is the noisy data at t, α_t is the scheduler which determines the noise variance at each step, and $\mathbf{I}$ is the identity matrix.

The reverse process, which is a learned neural network parameterized by θ, endeavours to reconstruct the original data from its noisy version, thus performing denoising:

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t)), \tag{5}$$

where the parameters of mean and variance are μ_θ and Σ_θ.

As proven in Ho et al. [10], x_{t-1} can be computed from x_t:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(x_t,t)) + \sigma_t \mathbf{z} \tag{6}$$

where $\mathbf{z} \sim \mathcal{N}(0,\mathbf{I})$, $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$, ε_θ is a neural network function that learns noise prediction, and σ_t is the standard deviation schedule.

We inform our model with the input image and its dynamically updated cumulative mask at each depth layer. This allows the model to recover the occluded objects progressively based on previously learned context. We achieve this by concatenating a given image I, the cumulative mask CM_i and amodal mask M_i for objects in layer L_i along the channel dimension, and define:

$$\mathbf{X_i} := I \oplus \mathrm{CM}_i \oplus M_i \tag{7}$$

The forward processing of q adds noise only to the amodal masks, keeping the input image and the corresponding cumulative mask unaltered. For a given image I and cumulative mask CM_i, we only add noise to the amodal mask M_i:

$$M_{i,t} = \sqrt{\bar{\alpha}_t}M_i + \sqrt{1-\bar{\alpha}_t}\varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\mathbf{I}) \tag{8}$$

Since we can define $\mathbf{X_{i,t}} := I \oplus \mathrm{CM}_i \oplus M_{i,t}$, Equation 6 is modified as,

$$M_{i,t-1} = \frac{1}{\sqrt{\alpha_t}}(M_{i,t} - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_\theta(X_{i,t},t)) + \sigma_t \mathbf{z} \tag{9}$$

where $\mathbf{z} \sim \mathcal{N}(0,\mathbf{I})$. The reverse process aims to reconstruct the noise-free amodal mask from its noisy counterpart, effectively denoising the mask at each timestep as t decreases.

The neural network's parameters are trained to minimize the difference, measured by the Kullback-Leibler divergence, between the forward and reverse distributions across all timesteps. The loss function is expressed as:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\mathbf{M}_i,\varepsilon}\left[\|\varepsilon - \varepsilon_\theta(\sqrt{\bar{\alpha}_t}M_i + \sqrt{1-\bar{\alpha}_t}\varepsilon, t)\|^2\right], \tag{10}$$

where ε is the true noise, and ε_θ is the model-predicted noise. The training process optimizes θ by minimizing the mean squared error between the true and predicted noise, facilitating a precise recovery of the amodal mask through the reverse diffusion sequence.

During inference, the model utilizes the learned reversal mechanism to generate multiple plausible amodal masks by sampling from a standard Gaussian distribution and conditioning on each object's unique context:

$$M_{gen,i}^{(k)} = f_\theta(\mathcal{N}(0,\mathbf{I}), I, \mathrm{CM}_i), \quad k = 1,\ldots,K, \tag{11}$$

where f_θ represents the trained generative function of the model, and $M_{gen,i}^{(k)}$ is the k-th generated amodal mask prediction for the object O_i. This process allows the generation of multiple plausible occlusion masks for each object layer.

4.2 Cumulative Occlusion Learning

Lack of spatial contextual awareness of surrounding objects in amodal segmentation can yield inaccurate or incomplete scene interpretations. To address this, we propose the cumulative occlusion learning algorithm, which employs a hierarchical procedure that learns to predict amodal masks in an order-aware manner. It operates by accumulating visual information, where the history of observed data (previous segmentation masks) influences the perception of the current data (the current object to segment). This strategy is akin to human perception, where the understanding of a scene is constructed incrementally as each object is identified and its spatial relation to others is established.

Training. We initiate with an empty cumulative mask (CM_0) and an image I with N layers. The model proceeds iteratively, predicting the amodal mask $\hat{M}_i$ for each layer while updating the cumulative mask using ground truth amodal masks to ensure the accuracy of the spatial context during training. Note that the diffusion is applied solely to the amodal mask predictions, while the image I and the cumulative mask CM remain intact. This cumulative strategy enhances accuracy by incorporating occlusion context into each layer in the learning process, enhancing the model's spatial understanding. Alg. 1 shows the complete training process. Notably, we introduce a predictive step for a layer $N+1$, which trains the model to expect a blank mask after all object layers have been identified and segmented. This ensures that the model learns to identify the last layer with any partially-visible objects and does not continue to hallucinate fully-occluded objects behind these.

Algorithm 1 Training Algorithm for cumulative occlusion learning

---

Input: Image I with number of N layers
Output: Ordered sequence of amodal masks $\tau = \langle \hat{M}_1, \hat{M}_2, \ldots, \hat{M}_N \rangle$
Initialize CM_0 to a blank mask; Initialize the ordered sequence τ as an empty list
for $i = 1$ to N **do**
    Input to model: I, CM_{i-1}; Predict amodal mask $\hat{M}_i$ for objects in layer L_i; Update $\mathrm{CM}_i \leftarrow \mathrm{CM}_{i-1} \cup M_i$ (Ground Truth); Append $\hat{M}_i$ to the sequence τ
end for
Perform a final prediction $\hat{M}_{N+1}$ with I and CM_N
assert $\hat{M}_{N+1}$ is a blank mask
return τ

Inference. Different from training, the inference phase needs to operate without available ground truth. Thus, it selects the most representative amodal mask from multiple predictions generated by the diffusion model to update the cumulative mask. Inference commences with an image I and aims to reconstruct an ordered sequence of amodal masks by layer. For each layer, a set of K diffusion-generated amodal mask predictions are evaluated to select the most representative amodal mask $\hat{M}_i$ for that layer. The selection criterion is based on the minimum absolute difference from each mask to the mean of non-null predictions, while ensuring spatial continuity between consecutive layers. The selected mask is then utilized to update the cumulative mask for subsequent layers' predictions. The process continues iteratively for an image I until a stopping criterion is met. The stopping criteria are established to avoid over-generation of invalid predictions when (1) reaching the maximum number of layers, or (2) all predicted masks are empty or the predicted object pixels of the selected mask are below a threshold area.

Alg. 2 shows the complete inference process, where the stopping criteria N_{max} and $Area_{min}$ are determined by the maximum number of layers and the minimum object area present in the corresponding training data, respectively. The appendix discusses the difference between training and inference in more detail.

Algorithm 2 Inference Algorithm for cumulative occlusion learning

Input: Image I, Maximum number of layers N_{max}, Minimum object pixel area $Area_{min}$
Output: Ordered sequence of amodal masks $\tau = \langle \hat{M}_1, \hat{M}_2, \ldots \rangle$
Initialize CM_0 to a blank mask; Initialize the ordered sequence τ as an empty list; Initialize $i = 1$
while $i \leq N_{max}$ **do**
  Generate K mask predictions $\{\hat{M}_i^1, \hat{M}_i^2, \ldots, \hat{M}_i^K\}$; Compute mean map $\overline{M}_i$ from non-null $\hat{M}_i^j$; Select $\hat{M}_i$ with minimum $||\hat{M}_i^k - \overline{M}_i||$; Enforce spatial integrity: if $\hat{M}_i \cap \hat{M}_{i-1} = \varnothing$, reassign $\hat{M}_i$ to the same layer as $\hat{M}_{i-1}$
  if $\hat{M}_i$ is null or $\hat{M}_i.area < Area_{min}$ **then** Break
  end if
  Update $CM_i \leftarrow CM_{i-1} \cup \hat{M}_i$; Append $\hat{M}_i$ to τ; $i \leftarrow i+1$
end while
return τ

In summary, cumulative occlusion learning allows the network to learn a robust internal representation of class-agnostic amodal object shape through occlusion, and to recognise the depth layer ordering of objects in scenes. This approach means the model can handle any arbitrary number of layers of occlusions, because it automatically learns to recognise when all visible objects have been segmented. Moreover, by preserving the input image and cumulative mask unaltered during the diffusion perturbations, our model maintains the fidelity of the contextual information, which is crucial for generating accurate amodal predictions.

5 Experiments

Datasets. We focus on amodal datasets highly relevant to robotics applications. Intra-AFruit, ACOM and MUVA [1, 14] include objects such as fruits, vegetables, groceries, and everyday products, effectively simulate the kind of visual clutter and occlusion challenges encountered in industrial robotics, making them ideal for our study. We enhanced these three datasets tailored for novel sequential amodal segmentation tasks, with layer structure annotations and class-agnostic masks. All images have been downsampled to a resolution of 64 × 64 pixels for computational efficiency. To eliminate indistinguishable or misleading ground truth data, we excluded images with post-downsampling visible object areas under 10 pixels. More details of the image processing are provided in the appendix.
Implementation Details. We set the timestep T=1,000 with a linear noise schedule for all the diffusion models. The models were trained using the AdamW optimizer [17] at a learning rate of 0.0001 and a batch size of 256. The other hyperparameters of the diffusion models follow the implementation in [20].
Evaluation metrics. The performance of class-agnostic segmentation is generally measured by comparing predicted masks with ground truth annotations [3, 23, 27]. We adopted two commonly used metrics: intersection over union (IOU) and average precision (AP).

5.1 Architecture Analysis

Number of generated amodal masks. Our method efficiently generates multiple amodal masks for each object (see Fig. 3a), capturing uncertainty and diverse occlusions without requiring varied annotations per image. This is particularly useful for amodal tasks considering occluded areas, where manual annotation is very expensive and synthetic images often provide only the sole ground truth.

While an arbitrary number of masks could be generated, we set to 3 masks per layer for inference in subsequent experiments, as this setting balances performance with computational demand on the ACOM dataset (see Tab. 1).

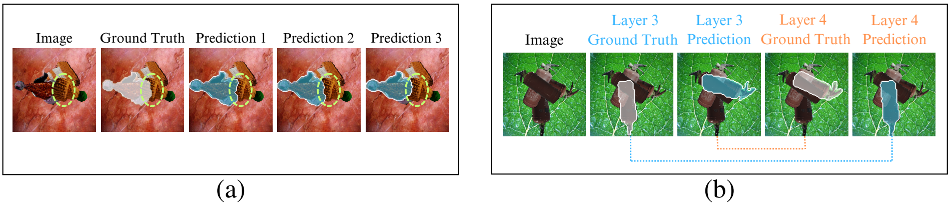

Figure 3: (a) Our approach considers the diversity of possible amodal masks, especially for occluded regions (indicated by dashed circles). (b) Example of misjudgement of the order of occluded objects in adjacent layers. Layer 3's prediction reflects Layer 4's ground truth and vice versa. This can also be a challenge for human perception.

Metric	Layer 1		Layer 2		Layer 3		Layer 4		Layer 5	
Ensemble	IOU	AP	IOU	AP	IOU	AP	IOU	AP	IOU	AP
k=3	**57.1**	**57.8**	**44.8**	**45.4**	28.8	30.0	12.2	14.2	1.9	3.6
k=5	56.7	57.5	44.3	44.9	28.8	29.7	12.7	**14.3**	2.3	**3.7**
k=7	56.8	57.5	44.7	45.4	29.4	30.0	12.6	14.1	2.6	3.6
k=9	56.9	57.7	44.4	45.1	**29.5**	**30.2**	**12.9**	14.2	2.4	3.7

Table 1: Ablation study for generating different numbers of masks during inference.

Selection of cumulative mask. The inference process could give multiple predictions for each layer, so there might be two options to update the cumulative mask for a given layer: (1) use one most representative prediction for that layer. We select the prediction with the minimum absolute difference from the mean of all predictions as the one. (2) use the mean of all predictions for that layer to form a mean mask. While the mean mask more explicitly takes into account all predictions, the risk is that when a prediction incorrectly gives an object that does not belong in that layer, the mean mask reacts to that as well. For example, a previous prediction showing an object in the next layer may cause the next prediction to ignore that object, because the object is already included in the given mean mask.

Therefore, in the inference process, the cumulative mask employs the most representative amodal mask (with the minimum absolute difference from the mean mask) rather than directly using the mean mask of all predictions for that layer. This avoids confusion due to the simultaneous prediction of objects in different layers. Tab. 2 shows the superiority of our mask selection method over using the mean mask for occluded layers on ACOM dataset.

Choice of Cumulative Mask	L1	L2	L3	L4	L5
	AP				
Mean mask	57.7	43.1	27.9	10.4	2.8
Selective mask	57.8 (+0.1)	45.4 (**+2.3**)	30.0 (**+2.1**)	14.2 (**+3.8**)	3.6 (**+0.8**)

Table 2: For predicting occluded objects (Layer L>1), the mask we selected is more suitable for constructing cumulative masks than using the mean mask directly.

Failure analysis. A common challenge arises from errors in sequential prediction, particularly determining which of two objects is in front of the other when the overlapping region is occluded by a third object. This may lead to objects being predicted in incorrect layers, as illustrated in Fig. 3 (b). Synthetic images can amplify this challenge due to fewer spatial cues (such as height in the image plane or scene semantics) to disambiguate occluded object order. Our cumulative occlusion learning mitigates the impact of these errors by considering the cumulative mask for all preceding layers. We demonstrate the robustness of our method to such failures through additional noise introduction experiments in the appendix.

5.2 Comparisons with Other Methods

We benchmark against DIS [32], a leading diffusion-based segmentation method. For comparison, we trained distinct DIS models for each layer under the same iterations and evaluated the segmentation results separately for each layer. Tab. 3 comprehensively compares our method and the improved DIS across different layers on three amodal datasets. The performance of the MUVA dataset after five layers is omitted because the performance of both models approaches zero. The superiority of our method is particularly evident in deeper layers, where our method maintains reasonable performances, whereas DIS shows a marked decline, especially in the MUVA dataset. These results highlight the robustness of cumulative occlusion learning in handling layered occlusions across various datasets, particularly in more complex scenarios involving multiple layers of object occlusion. Moreover, our iterative approach adapts to variable layers and learns a representation that can generate masks for any layer, which means it has fewer computing demands than training separate models to predict each layer.

Due to the lack of class-agnostic amodal segmentation methods with layer perception, we compare against category-specific methods like PLIn for amodal segmentation with occlusion layer prediction [1], AISFormer for amodal segmentation without layer perception [30], and PointRend for modal segmentation [12]. We trained these comparison models using category-labelled amodal masks to meet their requirement for category-specific learn-

	Layer	1	2	3	4	5
Dataset	Method	IOU / AP	IOU / AP	IOU / AP	IOU / AP	IOU / AP
Intra-AFruit	DIS	89.5 / 90.7	81.6 / 82.6	52.4 / 52.6	9.8 / 12.4	0.5 / 2.0
	Ours	**94.3 / 94.7**	**87.4 / 88.2**	**76.2 / 77.3**	**26.7 / 27.6**	**7.2 / 7.4**
ACOM	DIS	31.6 / 34.8	26.6 / 28.7	1.6 / 10.2	0.2 / 6.0	0.1 / 2.5
	Ours	**57.1 / 57.8**	**44.8 / 45.4**	**28.8 / 30.0**	**12.2 / 14.2**	**1.9 / 3.6**
MUVA	DIS	68.2 / 71.5	19.3 / 27.3	0.1 / 8.6	0.2 / 3.4	0 / 0.5
	Ours	**77.0 / 79.3**	**48.7 / 51.2**	**25.4 / 27.8**	**8.5 / 9.9**	**1.0 / 1.1**

Table 3: Comparison with a diffusion-based segmentation model [32] without cumulative occlusion learning. Our method exhibits great improvement in complex, deeper-layer scenes.

Dataset			Intra-AFruit		ACOM		MUVA	
Method	Supervision	Framework	AP w/ Layer	AP w/o Layer	AP w/ Layer	AP w/o Layer	AP w/ Layer	AP w/o Layer
PointRend	Supervised	CNN-based	N/A	70.9	N/A	22.0	N/A	38.9
AISFormer	Supervised	Transformer-based	N/A	70.4	N/A	34.9	N/A	49.7
PLIn	Weakly supervised	CNN-based	42.2	78.9	3.9	17.0	16.3	47.3
Ours	Supervised	Diffusion-based	**84.6**	**92.6**	**45.4**	**65.5**	**53.1**	**55.7**

Table 4: Comparison with category-specific segmentation models. PointRend [12], AISFormer [30] and PLIn [1] are trained on category-specific data, whereas our models are trained using class-agnostic data. We evaluate the models by focusing solely on the segmentation quality, disregarding any category information.

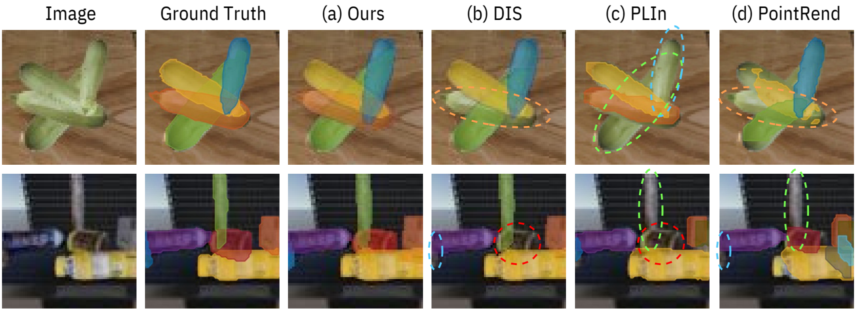

Figure 4: Comparison of predictions on Intra-AFruit (top) and MUVA (bottom) test image by (b) DIS [32] (c) PLIn [1] (d) PointRend [12] and (a) ours, where (b) and (c) are diffusion-based methods. Dashed circles indicate objects that missed being predicted. Others fail to segment objects or provide less plausible amodal masks compared to ours.

ing, while our model is trained on data without category labels. For evaluation, we ignore category label accuracy for the comparison models, reporting only segmentation accuracy.

We present the AP results considering two scenarios in Tab. 4: with layer prediction, where segmentation precision is contingent on correct layer assignment, and without layer prediction, where segmentation is recognized irrespective of layer placement. Despite being trained on class-agnostic data, our method surpasses category-specific models trained on category-labelled data. Furthermore, Fig. 4 visually demonstrates our method's superiority in amodal mask segmentation. Our approach provides plausible masks even for heavily-occluded objects, showcasing its enhanced segmentation capability in complex scenes involving multiple layers of object occlusion.

6 Limitation

Our method, while promising for sequential amodal segmentation, faces slow training and inference speeds due to the inherently computationally intensive nature of diffusion models. Diffusion models generally require a compression step for computational efficiency, and our current downsampling approach serves as an effective initial strategy. Future work will aim to augment efficiency and maintain output quality through super-resolution techniques and learned compression methods like VAEs, thus extending sequential amodal segmentation to high-resolution datasets such as KINS [21] and COCOA [39].

7 Conclusion

The task of sequential amodal segmentation is essential for understanding complex visual scenes where objects are frequently occluded. Our proposed method, leveraging cumulative occlusion learning with mask generation based on diffusion models, allows robust occlusion perception and amodal object segmentation over arbitrary numbers of occlusion layers. We demonstrate in three publicly-available amodal datasets that the proposed method outperforms other layer-perception amodal segmentation and diffusion segmentation methods while producing reasonably diverse results.

References

[1] Jiayang Ao, Qiuhong Ke, and Krista A Ehinger. Amodal intra-class instance segmentation: Synthetic datasets and benchmark. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 281–290, 2024.

[2] Seunghyeok Back, Joosoon Lee, Taewon Kim, Sangjun Noh, Raeyoung Kang, Seongho Bak, and Kyoobin Lee. Unseen object amodal instance segmentation via hierarchical occlusion modeling. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 5085–5092. IEEE, 2022.

[3] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8890–8899, 2020.

[4] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[5] 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)*, pages 6144–6153, 2018.

[6] 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.

[7] Kui Fu, Xuanju Dang, and Yichen Zhang. Taylor neural network for unseen object instance segmentation in hierarchical grasping. *IEEE/ASME Transactions on Mechatronics*, 2024.

[8] Jianxiong Gao, Xuelin Qian, Yikai Wang, Tianjun Xiao, Tong He, Zheng Zhang, and Yanwei Fu. Coarse-to-fine amodal segmentation with shape prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1262–1271, 2023.

[9] Zhitong Gao, Yucong Chen, Chuyu Zhang, and Xuming He. Modeling multimodal aleatoric uncertainty in segmentation with mixture of stochastic experts. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems (NeurIPS)*, 33:6840–6851, 2020.

[11] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 4019–4028, 2021.

[12] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 9799–9808, 2020.

[13] Hyunmin Lee and Jaesik Park. Instance-wise occlusion and depth orders in natural scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21210–21221, 2022.

[14] Zhixuan Li, Weining Ye, Juan Terven, Zachary Bennett, Ying Zheng, Tingting Jiang, and Tiejun Huang. Muva: A new large-scale benchmark for multi-view amodal instance segmentation in the shopping scenario. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23504–23513, October 2023.

[15] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7667–7676, October 2023.

[16] Zhaochen Liu, Zhixuan Li, and Tingting Jiang. Blade: Box-level supervised amodal segmentation through directed expansion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3846–3854, Mar. 2024.

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[18] Rohit Mohan and Abhinav Valada. Amodal panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21023–21032, 2022.

[19] Khoi Nguyen and Sinisa Todorovic. A weakly supervised amodal segmenter with boundary uncertainty estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7376–7385, 2021.

[20] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, pages 8162–8171. PMLR, 2021.

[21] 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*, pages 3014–3023, 2019.

[22] Lu Qi, Jason Kuen, Zhe Lin, Jiuxiang Gu, Fengyun Rao, Dian Li, Weidong Guo, Zhen Wen, Ming-Hsuan Yang, and Jiaya Jia. Ca-ssl: Class-agnostic semi-supervised learning for detection and segmentation. In *European Conference on Computer Vision (ECCV)*, pages 59–77. Springer, 2022.

[23] Ahmed Radwan and Mohamed S Shehata. Distilling part-whole hierarchical knowledge from a huge pretrained class agnostic segmentation framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 238–246, 2023.

[24] Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11536–11546, 2023.

[25] N Dinesh Reddy, Robert Tamburo, and Srinivasa G. Narasimhan. Walt: Watch and learn 2d amodal representation from time-lapse imagery. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9346–9356, 2022.

[26] Ronald A Rensink and James T Enns. Early completion of occluded objects. *Vision research*, 38(15-16):2489–2505, 1998.

[27] Mennatullah Siam, Alex Kendall, and Martin Jagersand. Video class agnostic segmentation benchmark for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2825–2834, 2021.

[28] Yihong Sun, Adam Kortylewski, and Alan Yuille. Amodal segmentation through out-of-task and out-of-distribution generalization with a bayesian model. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1205–1214, 2022.

[29] Hongkun Tian, Kechen Song, Song Li, Shuai Ma, Jing Xu, and Yunhui Yan. Data-driven robotic visual grasping detection for unknown objects: A problem-oriented review. *Expert Systems with Applications*, 211:118624, 2023.

[30] Minh Q Tran, Khoa HV Vo, Kashu Yamazaki, Arthur Fernandes, Michael T Kidd, and Ngan Le. Aisformer: Amodal instance segmentation with transformer. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24*, 2022.

[31] Kentaro Wada, Kei Okada, and Masayuki Inaba. Joint learning of instance and semantic segmentation for robotic pick-and-place with heavy occlusions in clutter. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9558–9564, 2019.

[32] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C Cattin. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning (MIDL)*, pages 1336–1348. PMLR, 2022.

[33] Junde Wu, RAO FU, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning (MIDL)*, 2023.

[34] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2955–2966, June 2023.

[35] Xiaoding Yuan, Adam Kortylewski, Yihong Sun, and Alan Yuille. Robust instance segmentation through reasoning about multi-object occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11141–11150, 2021.

[36] Lukas Zbinden, Lars Doorenbos, Theodoros Pissas, Adrian Thomas Huber, Raphael Sznitman, and Pablo Márquez-Neila. Stochastic segmentation with conditional categorical diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1119–1129, 2023.

[37] 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)*, pages 3784–3792, 2020.

[38] Chuanxia Zheng, Duy-Son Dao, Guoxian Song, Tat-Jen Cham, and Jianfei Cai. Visiting the invisible: Layer-by-layer completed scene decomposition. *International Journal of Computer Vision (IJCV)*, 129:3195–3215, 2021.

[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 (CVPR)*, pages 1464–1472, 2017.