

# Sequential Amodal Segmentation via Cumulative Occlusion Learning (Supplementary Material)

Jiayang Ao<sup>1</sup>  
jiayang.ao@student.unimelb.edu.au

QiuHong Ke<sup>2</sup>  
qiuHong.ke@monash.edu

Krista A. Ehinger<sup>1</sup>  
kehinger@unimelb.edu.au

<sup>1</sup> The University of Melbourne  
Parkville, VIC, Australia

<sup>2</sup> Monash University  
Clayton, VIC, Australia

---

In this supplementary material, we provided additional details for our work: we first provide the details of the image processing for the three enhanced amodal datasets. Then, we present the noise introduction experiment, which aligns the training and inference pipelines. Finally, we demonstrate more visualisations of the results for our model.

## A Datasets Processing

We experimented with three amodal datasets highly relevant to robotics applications in our paper, Intra-AFruit, ACOM and MUVA [1, 2]. We enhanced these three datasets tailored for the sequential amodal segmentation tasks, with layer structure annotations and class-agnostic masks. The training and test images in these datasets are sourced directly from the corresponding partitions of the original dataset. All images have been downsampled to a resolution of  $64 \times 64$  pixels for computational efficiency. Furthermore, we excluded images with post-downsampling visible object areas under 10 pixels to eliminate indistinguishable or misleading ground truth data.

**Intra-AFruit** [1] dataset contains ten classes of fruits and vegetables. We limited the original test set to a random subset of 3,000 images to enhance experimental efficiency. The reprocessed dataset includes 187,204 training and 3,000 test images, with each image potentially containing up to five layers.

**ACOM** [2] dataset contains ten classes of common objects with synthetically generated annotations. The reprocessed dataset includes 9,378 training and 2,355 test images with up to five layers.

**MUVA** [2] dataset contains twenty categories of supermarket items. To avoid compression distortion of non-square images, we cropped square images using the shortest edge and aligned the crop to the leftmost or centre, which follows object distribution rules to preserve more objects. The reprocessed dataset includes 5,582 training and 1,722 test images with up to seven layers.

## B Noise Introduction Experiment in Cumulative Mask

Our model leverages the ground truth cumulative mask as input during training, while inference uses the predicted masks from previous layers to build the cumulative mask. A common idea is to utilize the predicted cumulative mask in training, mirroring the inference setup. However, this complicates the early stages of training, when all of the predicted masks (and thus the cumulative mask) are similar to random noise. We conducted experiments in which we introduced controlled noise into the cumulative mask during training, to simulate the types of errors which occur during inference, but the results showed that this did not noticeably change the trained model’s performance. Therefore, the model presented in the main paper uses the ground truth cumulative mask during training.

This section details the experiments conducted to explore the impact of noise introduced during model training. As described in Sec. 4.2 of the main paper, we use ground truth cumulative masks during training, while utilising predicted cumulative masks during inference. To bridge the gap between training and inference, we introduce controlled noise into the cumulative mask during training.

In addition, the experiments simulate and seek to understand the impact of sequential prediction errors on the model’s performance, as discussed in the ‘Failure analysis’ of Sec. 5.1 in the main paper. By introducing noise into the cumulative mask during training, we effectively create scenarios where the model must handle instances segmented into the wrong layer, as happens when the model makes sequential prediction errors.

### B.1 Experimental Design

The experiment was designed to mimic common types of inference errors, such as continuous prediction errors due to layer dependencies or over-segmentation due to boundary ambiguity. This was achieved by selectively omitting instances from a random layer in the cumulative mask while keeping the input RGB image and the prediction mask unchanged.

Specifically, instances from a randomly chosen layer (excluding the fully visible layer) are excluded from the cumulative mask. Mathematically, selecting a random layer index  $i_{\text{rand}}$  from  $[2, n]$ , the perturbed version of the cumulative mask, denoted as  $P$ , is derived by:

$$P = CM - M_{i_{\text{rand}}} \quad (1)$$

Where  $CM$  is the original cumulative mask, and  $M_i$  is the ground truth mask of the  $i^{\text{th}}$  layer instance ( $i \in [2, n]$ ). The subtraction here is a pixel-wise binary operation. During training, the model will replace  $CM$  with  $P$  as input at a specified noise level ratio.

### B.2 Results and Analysis

The following experimental results show the model’s performance under varying noise levels, evaluated using Average Precision (AP) and Intersection Over Union (IOU) metrics. Tab. 1 illustrates the model’s performance in terms of AP and IOU across different layers and noise levels. It was observed that the highest AP was achieved with 0% noise for all layers. Similar to AP, the IOU results also showed that the highest performance was generally observed with 0% noise, except for the 5th layer, where a slight increase was noted at 10% noise level. Overall, this suggests that adding noise in training has very limited benefit. On the contrary, training without noise achieves the best performance in terms of AP or IOU in the vast majority of cases.

Noise	0%	5%	10%	15%	20%	0%	5%	10%	15%	20%
Layer	AP					IOU				
1	<b>57.8</b>	51.7	56.6	56.0	57.6	<b>57.1</b>	50.3	55.8	55.3	56.9
2	<b>45.4</b>	37.5	44.1	40.2	40.3	<b>44.8</b>	35.5	43.2	38.8	39.2
3	<b>30.0</b>	24.6	28.0	24.9	23.5	<b>28.8</b>	21.9	26.8	22.4	20.8
4	<b>14.2</b>	10.7	12.1	10.3	9.2	<b>12.2</b>	7.9	10.3	8.0	6.5
5	<b>3.6</b>	3.3	3.4	3.2	2.9	1.9	1.9	<b>2.2</b>	1.7	1.0

Table 1: Comparison at different noise levels, evaluated with AP and IOU. Noise-free training results in the highest AP across the layers, and the highest IOU for the first four layers and the second highest for the fifth layer.

The results of the experiment provide insight into the model’s robustness to errors in the sequential segmentation process and validate the effectiveness of our cumulative occlusion learning approach. By focusing on the cumulative mask for all preceding layers, our approach avoids the cascading effects of sequential prediction errors, ensuring more reliable performance even in complex occlusion scenarios.

Despite the theoretical appeal of mimicking inference conditions during training, the results indicate that using ground truth cumulative masks remains the more effective approach. This strategy consistently yielded superior results across most metrics and layers, showing its suitability to our model training process. Based on these findings, our training strategy uses the ground truth cumulative masks.

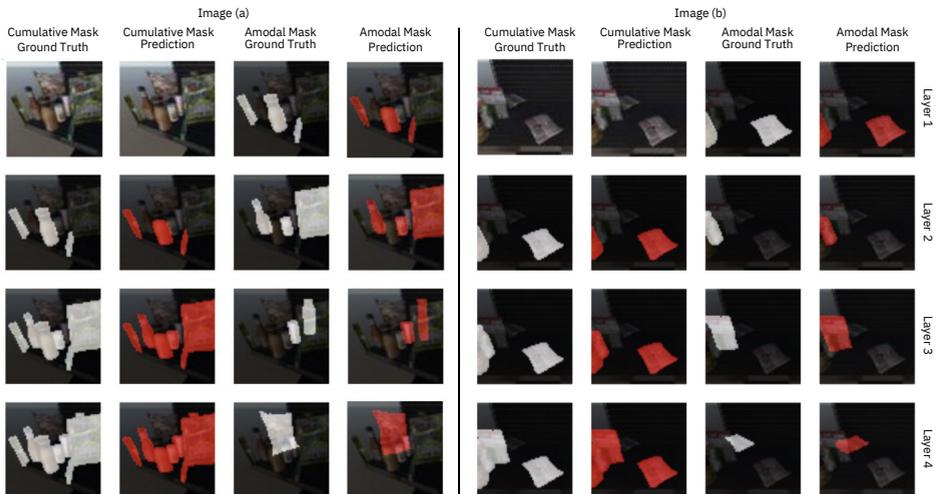


Figure 1: Visualisation of the prediction of our model on the MUVA [10] test set. Each layer’s amodal mask synthesis receives the cumulative mask of the previous layers as input, thus providing a spatial context for the prediction and helping to segment the remaining occluded objects better. We can see that our model can predict amodal masks and occlusion layers well for multiple objects in a given image.

## C More Visualisations for Our Results

We provide more visualisations of our model’s predictions for the MUVA [10] (Fig. 1), ACOM [10] (Fig. 2) and Intra-AFruit [10] (Fig. 3) test sets. As we can see from the figures, our model performs robustly with different objects and different levels of occlusion.

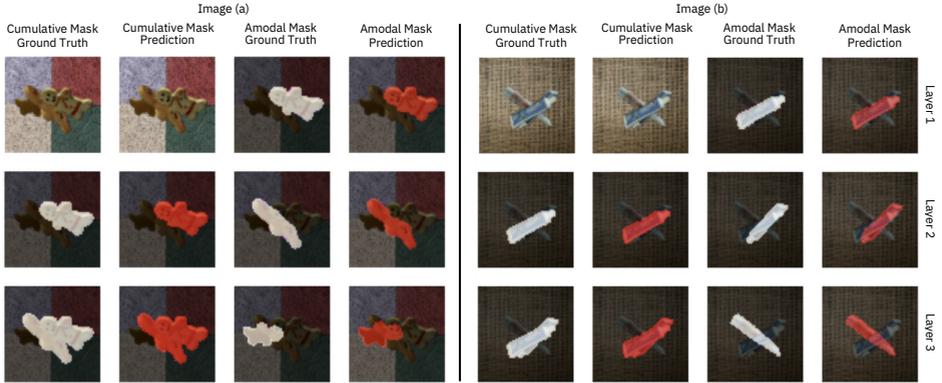


Figure 2: Visualisation of the prediction of our model on the ACOM [10] test set.

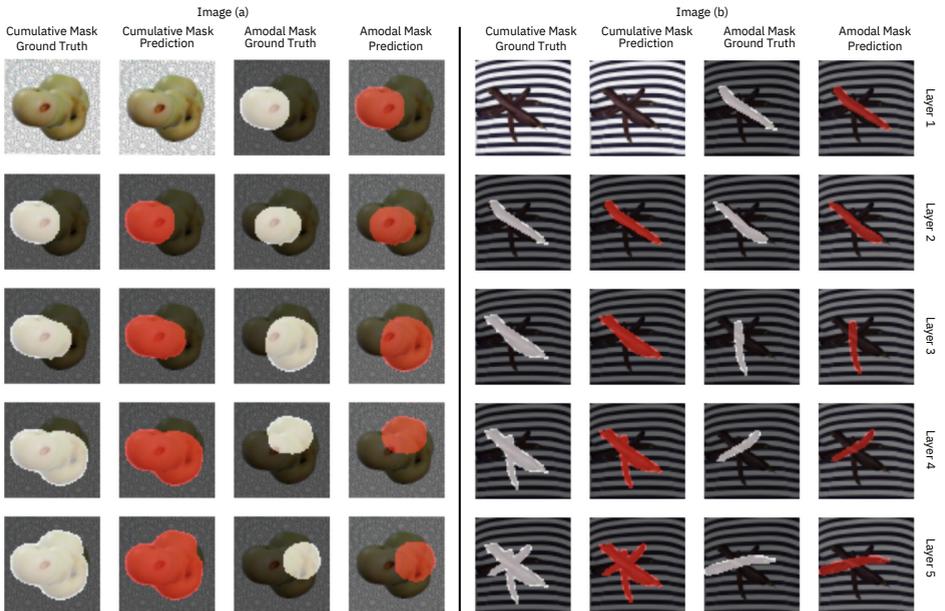


Figure 3: Visualisation of the prediction of our model on the Intra-AFruit [10] test set.

## References

- [1] Jiayang Ao, Qihong 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] 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.