Segment Anything, Even Occluded

Wei-En Tai[1*] Yu-Lin Shih[1*] Cheng Sun[2] Yu-Chiang Frank Wang[2,3] Hwann-Tzong Chen[1,4]
[1]National Tsing Hua University [2]NVIDIA [3]National Taiwan University [4]Aeolus Robotics

Abstract

Amodal instance segmentation, which aims to detect and segment both visible and invisible parts of objects in images, plays a crucial role in various applications, including autonomous driving, robotic manipulation, and scene understanding. While existing methods require training both front-end detectors and mask decoders jointly, this approach lacks flexibility and fails to leverage the strengths of pre-existing modal detectors. To address this limitation, we propose SAMEO, a novel framework that adapts the Segment Anything Model (SAM) as a versatile mask decoder capable of interfacing with various front-end detectors to enable mask prediction even for partially occluded objects. Acknowledging the constraints of limited amodal segmentation datasets, we introduce Amodal-LVIS, a large-scale synthetic dataset comprising 300K images derived from the modal LVIS and LVVIS datasets. This dataset significantly expands the training data available for amodal segmentation research. Our experimental results demonstrate that our approach, when trained on the newly extended dataset, including Amodal-LVIS, achieves remarkable zero-shot performance on both COCOA-cls and D2SA benchmarks, highlighting its potential for generalization to unseen scenarios.

1. Introduction

Human visual perception extends beyond what is directly visible in a scene. We can naturally imagine and understand the complete shape of partially occluded objects through a combination of object recognition and prior knowledge about object categories. Even when correctly classifying objects is difficult, we can often infer the complete shape of partially visible objects by analyzing visible parts and reasoning about common occlusion patterns [9, 20, 24]. Amodal instance segmentation seeks to replicate this remarkable human capability by detecting and localizing objects in images and predicting their complete shapes, including both visible and occluded portions (Figure 1).

Figure 1. Amodal segmentation examples: The top row shows the original images. The middle row displays EfficientSAM predicted modal masks that only cover the visible parts of objects. The bottom row illustrates amodal masks that reveal the complete object shapes predicted by our method, SAMEO—a Segment Anything Model Even under Occlusion.

An effective approach to addressing amodal instance segmentation is to divide the task into two main components: object detection and mask segmentation. In recent years, significant advances have been made in object detection, with state-of-the-art models such as RTMDet [16] and ConvNeXt-V2 [28] achieving impressive performance. However, current amodal segmentation approaches often require training both the detector and mask decoder jointly, which prevents them from fully utilizing these powerful pre-trained modal detectors. This limitation motivated us to develop a more flexible framework that can leverage existing modal detectors while still maintaining strong amodal segmentation capabilities. The emergence of foundation models for visual understanding has opened up new possibilities in segmentation tasks. Among these, the Segment Anything Model (SAM) [11] and its efficient variant, EfficientSAM [29], have demonstrated remarkable capabilities in prompt-based modal segmentation. We leverage EfficientSAM's architecture, which features a lightweight encoder for faster inference, and adapt it for amodal segmen-

*These authors contributed equally to this work.

tation through specialized training. Our approach enables the model to process both amodal and modal prompts for generating amodal mask predictions while maintaining potential zero-shot capabilities. Besides the improvements in algorithms and architectures, datasets are also crucial for learning-based methods, yet current amodal segmentation datasets encounter several challenges:

- **Limited Scale:** Existing datasets contain relatively few images, hindering the development of robust models.
- **Annotation Quality:** Several datasets relying on automatic generation methods can lead to inconsistent and sometimes incorrect instance annotations when not properly validated.
- **Irrelevant Objects:** A significant portion of annotated objects, such as walls and floors, contribute little to meaningful scene understanding.

To address these limitations, we present Amodal-LVIS, a new large-scale dataset derived from LVIS [6] and LVVIS [26]. Our dataset contains 300K carefully curated images, where each image contains one instance annotation. These annotations form paired examples between synthetic occluded instances and their original unoccluded versions. Furthermore, we have processed and refined existing datasets to create a comprehensive training collection comprising approximately 1M images and 2M instance annotations.

Experimental results show that our method with the EfficientSAM architecture, when trained on our combined dataset, achieves remarkable zero-shot performance that surpasses previous supervised amodal segmentation methods. These results validate our approach of leveraging an efficient existing architecture with high-quality, large-scale training data for amodal segmentation tasks.

Our main contributions can be summarized as follows:

1. **Flexible Amodal Framework:** The proposed method, SAMEO, adapts EfficientSAM for amodal instance segmentation that works with both modal and amodal detector prompts through specialized training.
2. **Large-scale Dataset:** A new Amodal-LVIS dataset containing 300K images, forming paired examples between synthetic occluded instances and their original unoccluded versions.
3. **Dataset Collection:** A comprehensive training collection of 1M images and 2M instances created by combining and refining existing amodal datasets with Amodal-LVIS.
4. **Zero-shot Performance:** State-of-the-art zero-shot results on both COCOA-cls and D2SA benchmarks, surpassing previous supervised methods.

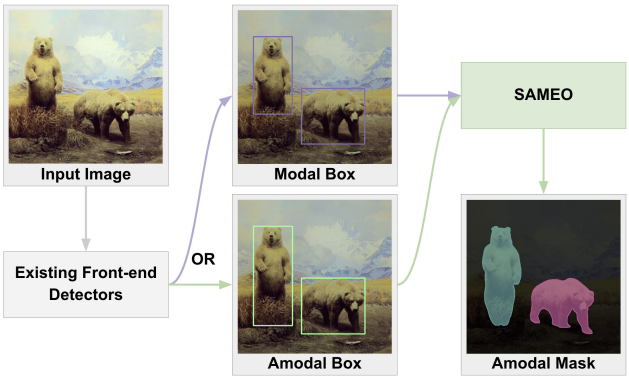

Figure 2. Overview of our amodal segmentation pipeline. Given an input image, existing object detectors first generate either modal boxes (showing visible regions) or amodal boxes (showing complete object extent). Our SAMEO then processes these detections to produce amodal masks that recover the full shape of objects, including occluded parts.

2. Related Work

2.1. Instance Segmentation

Instance segmentation is a fundamental computer vision task that simultaneously addresses object detection and segmentation, aiming to both locate objects in a scene and generate precise mask predictions for each detected instance. Initially focusing on visible parts of objects (modal instance segmentation), this field continues to evolve with the emergence of deep learning architectures. State-of-the-art methods demonstrate improvements through transformer-based feature extraction [27], modernized convolutions [21, 28], and optimized speeds [16]. Further detection models built upon the DETR architecture [1] have achieved additional advances through specialized query selection mechanisms and training schemes [33, 35].

Building upon modal instance segmentation, amodal instance segmentation extends the task to predict complete object shapes, including occluded regions. This extension is first formalized by Li and Malik [12], leading to various architectural innovations [5, 30, 31]. Notable approaches include ORCNN [4], ASN [22], which enhance Mask R-CNN [7] with occlusion reasoning capabilities, and BCNet [10] with its bilateral layers for handling object overlaps. Currently, AISFormer [25] represents the state-of-the-art in amodal instance segmentation by introducing transformers to effectively model long-range dependencies.

2.2. Segment Anything Model

The Segment Anything Model (SAM) [11] represents a significant advancement in foundational computer vision models, capable of segmenting any visual object based on various prompts, including points or boxes. Trained on a dataset

of 11M images, SAM demonstrates outstanding zero-shot generalization capabilities across diverse object categories and domains.

The original SAM model, despite its strong performance, faces practical limitations due to high computational demands, including significant memory requirements and slow inference speed. EfficientSAM [29] addresses these challenges by using a Masked Autoencoder (MAE) [8] pre-training method to learn the feature embeddings from SAM's original ViT-H encoder, resulting in faster inference speed and reduced model size while maintaining comparable segmentation performance.

2.3. Amodal Datasets

Several datasets have been introduced for amodal segmentation. COCOA [34] is the first amodal dataset, providing semantic-level amodal annotations for COCO images. D2SA/COCOA-cls [4] extends this with instance-level annotations. DYCE [3] offers synthetic indoor scenes with accurate ground truth. The KINS dataset [22] focuses on traffic scenarios with 14K images of vehicles and pedestrians. More recently, MUVA [13] introduces a multi-view shopping scenario dataset, while MP3D-Amodal [32] provides real-world indoor scenes from Matterport3D. WALT [23] uniquely utilizes time-lapse imagery to obtain amodal ground truth, and KITTI-360-APS [18] extends KITTI-360 [14] with amodal panoptic annotations. Furthermore, datasets from related amodal completion works, such as pix2gestalt [19], have contributed to the development of the field.

3. Our Approach

3.1. Enabling Amodal Mask Prediction

Preliminaries: EfficientSAM. Segment Anything Model (SAM) [11] is a foundation model for image segmentation that can generate high-quality object masks based on any prompt. The original SAM architecture consists of three main components: i) an image encoder that transforms the input image into image embeddings, ii) a lightweight transformer-based prompt encoder that converts prompts (points, boxes) into unified embeddings, and iii) a mask decoder that utilizes a transformer architecture with two cross-attention layers to process both image and prompt embeddings for generating the final segmentation mask. In our approach, we mainly use EfficientSAM [29], a compact adaptation of the original SAM model. EfficientSAM replaces SAM's image encoder with a lightweight ViT variant [2] while maintaining the original prompt encoder and mask decoder.

Model Architecture. We propose SAMEO for amodal instance segmentation, retaining a lightweight image encoder $\mathcal{E}$ as in the original architecture of EfficientSAM, a transformer-based prompt encoder $\mathcal{P}$, and a mask decoder $\mathcal{D}$ with dual cross-attention layers. Given an input image I and a bounding box prompt B, the proposed SAMEO pipeline predicts the amodal mask $\hat{M}$ and the estimated IoU $\hat{\rho}$ as follows:

$$\hat{M}, \hat{\rho} = \mathcal{D}(\mathcal{E}(I), \mathcal{P}(B)) \,. \tag{1}$$

Training Strategy. During training, we exclusively finetune EfficientSAM's mask decoder while keeping the original weights of the image encoder and prompt encoder unchanged. The model receives two inputs: an image and a bounding box prompt derived from ground-truth annotations. The box prompts are randomly selected from modal and amodal ground-truth boxes with equal probability. The training objective combines Dice loss [17], Focal loss [15], and L1 loss for IoU estimation:

$$\mathcal{L} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{Focal}} + \lambda \mathcal{L}_{\text{IoU}} \,, \tag{2}$$

where

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2|\hat{M} \cap M_{\text{gt}}|}{|\hat{M}| + |M_{\text{gt}}|} \,, \tag{3}$$

$$\mathcal{L}_{\text{Focal}} = -(1 - p_t)^{\gamma} \log(p_t) \,, \tag{4}$$

$$\mathcal{L}_{\text{IoU}} = \left| \hat{\rho} - \text{IoU}(\hat{M}, M_{\text{gt}}) \right| \,. \tag{5}$$

Here, M_{gt} represents the ground truth amodal mask, p_t is the predicted probability for the target class, γ is the focusing parameter set to 2 in our experiments, and λ is empirically set to 0.05.

Inference Pipeline. For inference, SAMEO can be flexibly integrated with various object detectors, including both amodal detectors (*e.g.*, AISFormer [25]) and conventional modal detectors (*e.g.*, RTMDet [16]). The detection outputs serve as box prompts for our model, which then generates corresponding amodal masks (Figure 2). This modular design allows our model to enhance existing detection systems with amodal segmentation capabilities while achieving state-of-the-art performance.

3.2. Amodal Dataset Collection

Limitations of Existing Amodal Datasets. Existing amodal datasets have inherent limitations in both manual and synthetic annotation mechanisms. Human-annotated datasets, while closely representing real-world scenarios, are costly to produce and prone to errors in occluded region estimation. Synthetic datasets, though efficient to generate, lack reliable verification mechanisms for object completeness and may not accurately reflect natural occlusion patterns (Figure 3).

Dataset	Type	# Instances	# Images	POI (%)	Average ROR (%)
COCOA (*no stuff*) [34]	Real	32,926	5,073	54.9	30.2
COCOA-cls [4]	Real	10,562	3,499	49	21.8
KINS [22]	Real	188,085	14,993	54.3	42.2
KITTI-360-APS [18]	Real	89,938	12,496	50	33.9
D2SA [4]	Synthetic	28,720	5,600	53.6	23.7
MUVA [13]	Synthetic	198,573	26,406	76.7	32.1
WALT* [23]	Synthetic	485,369	40,000	32	36.5
DYCE* [3]	Synthetic	66,453	5,229	77.8	29
MP3D-amodal* [32]	Synthetic	2,968	2,549	100	41.1
pix2gestalt [19]	Synthetic	849,667	849,667	100	35.9
Amodal-LVIS	Synthetic	399,398	301,493	50	34.5
Total	–	2,352,659	1,267,424	–	–

Table 1. Amodal dataset collection. Datasets marked with * are generated or refined by ourselves. POI is calculated as the percentage of occluded instances (where modal mask is not equal to amodal mask) over all instances. Meanwhile, average ROR is the average ratio of occluded regions, considering only instances with occlusions.

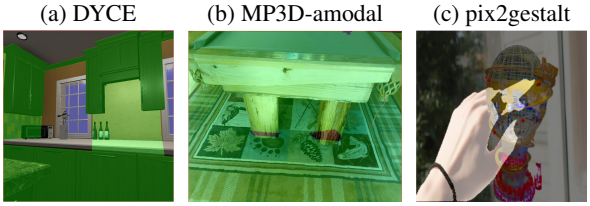

Figure 3. Examples of limitations in existing amodal datasets: (a) DYCE and (b) MP3D-amodal show meaningless architectural elements rendered from 3D meshes that dominate the image space, while (c) pix2gestalt contains potentially incomplete amodal masks due to restrictive generation criteria.

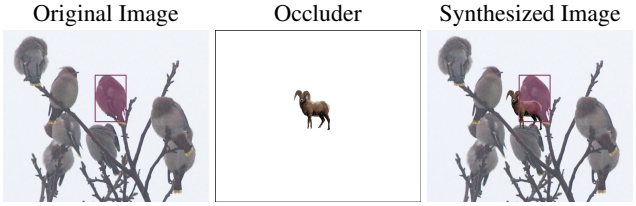

Figure 4. Amodal-LVIS dataset generation process. From left to right: original image with unoccluded objects, a selected occluder object, and the synthesized image with occlusion. Our dataset includes both the original and the synthesized image for each instance to prevent occlusion bias during training.

Dataset Collection and Quality Control. To leverage the advantages of both mechanisms, we have collected and filtered datasets of both annotation types. Our cleaning process addresses specific issues in each dataset to ensure data quality while maintaining realistic occlusion representations (Table 1).

For synthetic datasets DYCE and MP3D-amodal, generated using 3D furniture meshes, we identify and address two main quality issues: meaningless architectural elements (walls, floors, ceilings) occupying the majority of image space and objects with minimal visible areas. We implement filters to remove cases where visible parts are less than 10% of the whole object, objects occupying more than 90% of the image area, and architectural element annotations.

The WALT dataset leverages road surveillance time-lapse footage for synthetic data generation. It obtains bounding boxes for cars and people using a pre-trained detector and then identifies complete objects by analyzing these bounding box intersections. These discovered complete objects are then composited back into the same scenes to generate synthetic training data. However, their layer-by-layer placement can create unrealistic occlusions. We address this by implementing an occlusion threshold filter to ensure natural occlusion patterns.

For other datasets with class annotation, such as COCOA, the availability of semantic labels enabled straightforward quality control. We filter out "stuff" class annotations across these datasets to focus on meaningful objects that align with amodal instance segmentation goals.

3.3. Amodal-LVIS

We propose a synthetic dataset for amodal mask segmentation through precise object occlusion generation, incorporating complete object collection, synthetic occlusion generation, and a dual annotation mechanism to prevent model bias. Combined with existing datasets, our collection totals 1M images and 2M instance annotations.

Complete Object Collection. To obtain complete objects for synthetic occlusion, we utilize SAMEO, which is pre-trained on the previously mentioned amodal datasets, to generate pseudo labels for instances within LVIS and LVVIS datasets. Our model predicts amodal masks for each instance, which are then compared with the ground-truth visible mask annotations. This comparison helps us identify complete, unoccluded objects.

Synthetic Occlusion Generation. The occlusion generation process involves pairing randomly selected complete objects from our collected pool. To ensure that the occlusions look realistic, we normalize the paired objects to similar sizes while maintaining their aspect ratios. Object positioning and occlusion rates are controlled using bounding box annotations, which allow for precise management of how objects occlude one another.

Dual Annotation Mechanism. Our experiments in Section 4 show that training solely on occluded masks leads to model confusion, resulting in an over-prediction of occluded objects even when the prompts are intended to target foreground instances. To resolve this issue, we include both occluded and original unoccluded versions of instances in our dataset (Figure 4). This dual annotation mechanism prevents occlusion bias while providing comprehensive training examples for both states.

4. Experiments

4.1. Settings

Implementation Details. Our model is trained on NVIDIA Tesla V100/A100 GPUs for 1,440/2,340/22,500 iterations on COCOA-cls/D2SA/MUVA respectively. For zero-shot SAMEO, we increase the batch size to 32 and train for 40,000 iterations. We use the Adam optimizer with a learning rate of 1×10^{-4} without any learning rate scheduler. For each instance during training, we randomly select either the amodal or modal ground truth bounding box as the prompt with equal probability.

Datasets and Baselines. For training, we use our dataset collection and the proposed Amodal-LVIS dataset. As for evaluation, COCOA-cls, D2SA and MUVA datasets are utilized. We primarily compare our approach against AISFormer, the current state-of-the-art model in amodal instance segmentation. Unlike conventional instance segmentation models that incorporate both object detection and mask decoding components, SAMEO functions solely as a mask decoder. This allows our model to flexibly integrate with existing amodal and modal instance segmentation models, using their object box predictions as prompts to generate refined amodal masks. For a comprehensive evaluation, we compare our results against the original mask predictions from these front-end models.

To evaluate our zero-shot performance, we extend the comparisons beyond AISFormer to include both modal and amodal front-end models equipped with the original EfficientSAM as their mask decoder. This comparison demonstrates our successful adaptation of EfficientSAM for amodal mask segmentation while maintaining zero-shot capabilities.

Evaluation Metrics. We evaluate our method using two standard metrics: mean Average Precision (AP) and mean Average Recall (AR). Since our model is class-agnostic, we compute both metrics without considering class labels. For a fair comparison, we reproduce baseline methods and evaluate them using the same class-agnostic AP and AR. For methods that use SAMEO as mask decoder, we refine the confidence score $\hat{\rho}_{\text{front}}$ of front-end models using the estimated IoU $\hat{\rho}_{\text{ours}}$ predicted by SAMEO when calculating these metrics. The refined confidence score $\hat{\rho}_{\text{ref}}$ of these cases is computed as follows:

$$\hat{\rho}_{\text{ref}} = \hat{\rho}_{\text{front}} \times \hat{\rho}_{\text{ours}} \,. \tag{6}$$

4.2. Results

Quantitative Results. We evaluate our proposed SAMEO on three widely-used datasets: COCOA-cls, D2SA, and MUVA. For each dataset, we train our model on their respective training sets and evaluate on the corresponding test sets (Table 2). To demonstrate SAMEO's effectiveness and versatility, we attach it to various pre-trained front-end models, where modal front-ends are trained with modal annotations and amodal front-ends are trained with amodal annotations of these datasets. Our experimental results show that SAMEO significantly outperforms the current state-of-the-art method, AISFormer, achieving higher AP and AR across all datasets. Notably, our model exhibits robust performance in mask refinement independent of the front-end model's original mask type. Regardless of whether the front-end models produce modal or amodal mask predictions, SAMEO successfully refines them to achieve comparable high performance, demonstrating its strong capability to utilize both types of prompts.

Qualitative Results. For qualitative evaluation, we compare our model's predictions against AISFormer on both COCOA-cls and MUVA datasets (Figure 5). Our method exhibits superior performance across various challenging scenarios, including complex still life arrangements with multiple overlapping objects (*e.g.*, bottles and containers),

Model	COCOA-cls				D2SA				MUVA			
	AP	AP_{50}	AP_{75}	AR	AP	AP_{50}	AP_{75}	AR	AP	AP_{50}	AP_{75}	AR
AISFormer [25]	40.6	70.5	42.5	55.2	66.3	89.9	72.8	76.1	69.4	90.3	75.6	80.7
RTMDet* [16]	49.8	71.2	54.7	69.6	59.7	81.3	63.4	78.2	46.0	68.2	46.4	57.8
ConvNeXt-V2* [28]	46.7	67.5	52.0	70.9	66.1	89.1	69.8	75.3	47.9	69.0	46.6	61.7
ViTDet* [27]	47.4	69.4	52.5	68.8	63.0	86.4	65.8	75.4	45.3	66.9	44.7	60.7
AISFormer+SAMEO	54.3	74.0	59.7	69.3	79.8	92.7	84.8	84.2	76.2	90.8	80.9	**85.2**
RTMDet*+SAMEO	**55.3**	**75.2**	**60.8**	**74.3**	72.7	85.8	77.5	84.2	75.8	89.2	79.9	83.1
ConvNeXt-V2*+SAMEO	54.1	73.1	59.3	74.0	**80.8**	**94.0**	**85.1**	**87.1**	**79.2**	**93.1**	**82.6**	81.3
ViTDet*+SAMEO	54.1	73.3	59.2	72.3	78.6	92.3	82.6	84.7	74.1	89.0	78.2	82.3

Table 2. Quantitative comparison on various datasets. Models marked with * are modal instance segmentation methods that detect objects and segment their visible masks without handling occlusions, with performance metrics calculated using their modal mask predictions. SAMEO takes bounding box predictions from the front-end models as prompts to generate amodal instance masks. Evaluation metrics include Average Precision (AP) at different IoU thresholds and Average Recall (AR). Bold numbers indicate the best performance.

Model	COCOA-cls				D2SA			
	AP	AP_{50}	AP_{75}	AR	AP	AP_{50}	AP_{75}	AR
AISFormer	40.6	70.5	42.5	55.2	66.3	89.9	72.8	76.1
AISFormer+EfficientSAM†	47.6	70.0	51.7	64.2	69.6	89.2	72.3	77.7
RTMDet*+EfficientSAM†	48.7	71.1	53.2	65.9	63.0	82.7	64.8	74.3
RetinaNet* [21]+EfficientSAM†	44.8	67.1	48.6	68.4	60.6	77.5	62.9	79.3
DINO* [33]+SAMEO†	50.2	70.4	55.7	**73.7**	69.8	86.2	72.6	75.9
RetinaNet*+SAMEO†	51.0	72.2	56.6	72.6	62.5	78.0	64.8	80.7
AISFormer+SAMEO†	52.8	73.4	57.9	67.9	74.1	90.2	78.3	79.9
ViTDet*+SAMEO†	53.2	73.0	58.6	71.2	72.1	89.0	74.9	80.7
ConvNeXt-V2*+SAMEO†	53.4	73.0	59.0	73.0	74.3	**91.6**	78.0	**83.1**
CO-DETR* [35]+SAMEO†	54.0	74.8	60.2	73.5	**75.0**	91.0	**78.5**	82.2
RTMDet*+SAMEO†	**54.4**	**75.0**	**60.2**	73.4	68.4	84.5	71.0	77.7

Table 3. Zero-shot performance on COCOA-cls and D2SA datasets. The results show SAMEO not only significantly outperforms AISFormer but also successfully adapts EfficientSAM's modal segmentation capability to amodal segmentation, demonstrating consistent performance improvements when paired with various front-end detectors. † indicates zero-shot evaluation without training on the test dataset. * denotes modal object detectors that provide modal bounding boxes as prompts. Bold numbers indicate the best performance.

scenes with intricate occlusions (*e.g.*, people behind barriers), and diverse object categories and poses. Results show that our model generates significantly more precise amodal masks with sharper boundaries while providing more reasonable predictions for occluded parts. The qualitative comparison clearly demonstrates our method's improvements over the baseline method in both mask quality and occlusion reasoning capabilities, validating the effectiveness of SAMEO for real-world amodal segmentation tasks.

4.3. Zero-shot Performance

To evaluate SAMEO's zero-shot generalization capability, we train our model on our dataset collection and the proposed Amodal-LVIS dataset, excluding COCOA-cls and D2SA. During training, for each batch, a dataset is sampled with probability proportional to the logarithm of its size divided by the sum of log sizes across all datasets. We then test these two held-out datasets to demonstrate zero-shot performance (Table 3). For comparison, we include AISFormer and RTMDet (both trained on target datasets) combined with the original EfficientSAM, showing that our model successfully adapts EfficientSAM for amodal segmentation while preserving its zero-shot capability. Furthermore, we experiment with various pre-trained modal front-end detectors to demonstrate SAMEO's robust zero-shot performance regardless of the front-end choice.

The results demonstrate SAMEO's superior performance, achieving up to 13.8 AP improvement over AISFormer on COCOA-cls with RTMDet and 8.7 AP over on D2SA with CO-DETR, reaching state-of-the-art results and validating its strength as a robust zero-shot amodal segmentation solution.

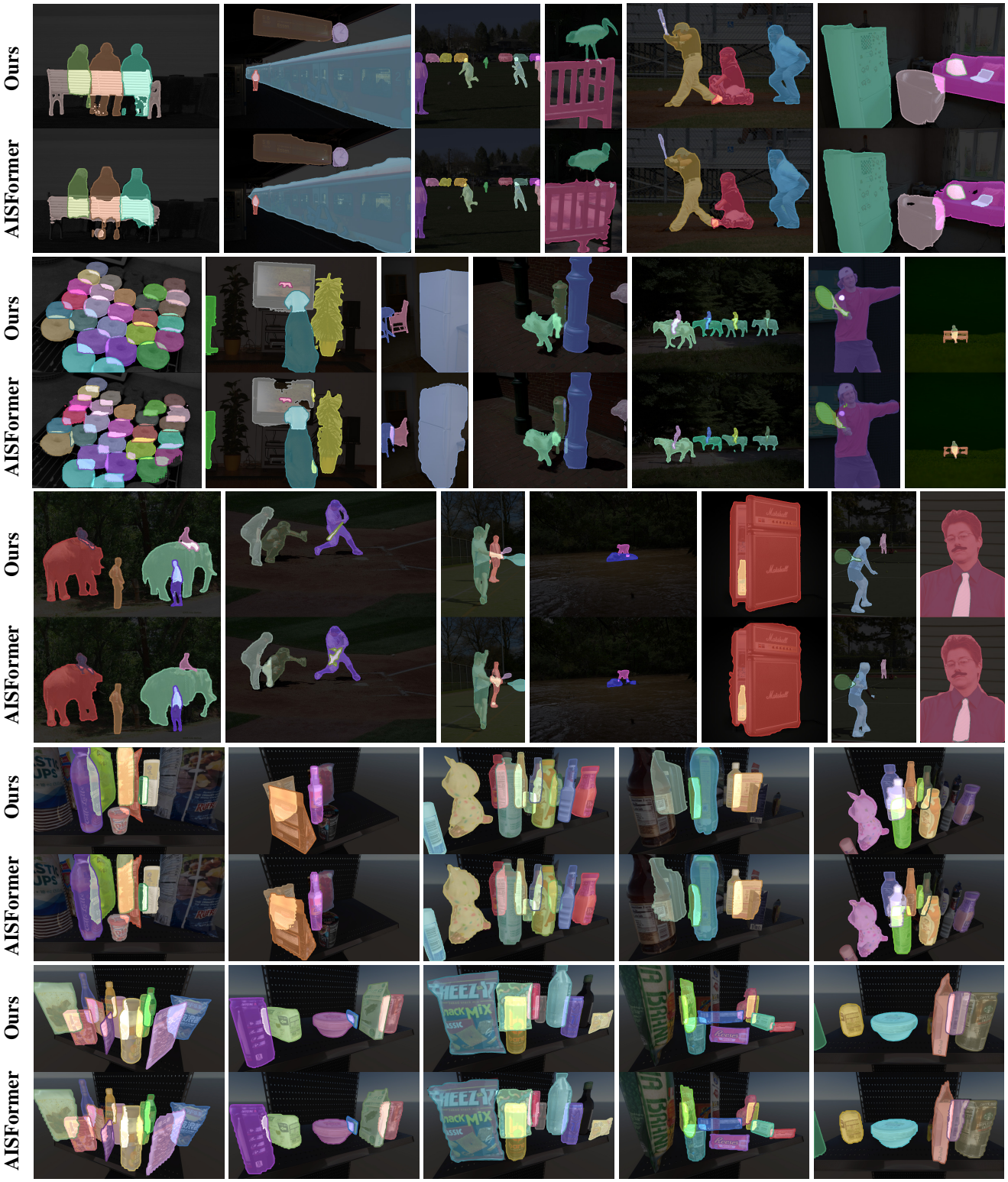

Figure 5. Qualitative comparison of amodal mask predictions. For each row: SAMEO's amodal prediction (top) with AISFormer box prompts, and AISFormer's prediction (bottom). Our method demonstrates superior mask quality, exhibiting more precise boundary delineation and robust handling of complex occlusion scenarios. Original images used for evaluation are available in supplementary materials.

IoU Prediction	COCOA-cls		
	AP	AP_{50}	AP_{75}
×	52.4	73.2	57.8
✓	**54.3**	**74.0**	**59.7**

Table 4. Ablation study of IoU prediction refinement on COCOA-cls dataset, using AISFormer as front-end. × and ✓ indicate without and with IoU prediction refinement, respectively.

Box Prompt	COCOA-cls			
	AP	AP_{50}	AP_{75}	AR
amodal	53.0	72.9	58.0	71.1
modal	53.7	73.3	59.3	71.2
random	**54.2**	**73.5**	**59.5**	**71.6**

Table 5. Comparison of different prompt types during training. We evaluate three variants on the COCOA-cls dataset using both modal and amodal front-ends, reporting the averaged performance over both types. The results show that training with random box prompts has optimal performance across diverse front-end models.

4.4. Ablation Study

Effect of IoU Prediction. To validate the effectiveness of the IoU prediction branch in our model, we have conducted experiments comparing the performance metrics before and after confidence score refinement using SAMEO's predicted IoU. Using AISFormer as our front-end model and evaluating this setting on the COCOA-cls dataset, the experimental results demonstrate that SAMEO's precise IoU prediction significantly contributes to improving the ranking of segmentation results (Table 4). Specifically, we observe that incorporating the predicted IoU for confidence score refinement leads to notable improvements in AP metric, confirming that the IoU prediction branch plays a crucial role in enhancing the overall performance of our model.

Impact of Training Prompt Types. We investigate the optimal prompt-type strategy for training SAMEO to achieve balanced performance across both modal and amodal front-end prompts. We train three variants of SAMEO using ground truth amodal boxes, modal boxes, and a random mixture of both with equal probability. For evaluation, we integrate each trained model variant with both amodal and modal front-end detectors and evaluate their performance separately. The final performance metric is calculated by averaging the AP and AR scores across both front-end scenarios (Table 5). Our findings reveal that training with random boxes with equal probability yields the most balanced performance when handling various front-end prompt types, demonstrating the model's ability to generalize across different input scenarios.

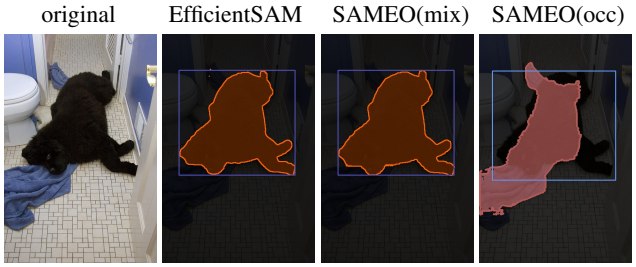

Figure 6. Visualization of over-prediction. Given a detector box clearly intended for the foreground dog, SAMEO(occ), trained solely on occluded instances, mistakenly predicts the mask of the background towel. After rebalancing our dataset to include both occluded and non-occluded instances, SAMEO(mix) produces results that closely match the EfficientSAM baseline.

Dataset Composition Analysis. To understand the importance of diverse instance types in training data, we have conducted an experiment training SAMEO exclusively on datasets containing only occluded instances (*e.g.*, pix2gestalt). Visualization results reveal a significant limitation: the model exhibits over-prediction of background instances, even when the input box prompt clearly indicates a foreground object (Figure 6). This observation motivates our design choice for the Amodal-LVIS dataset, which maintains an equal distribution of occluded and non-occluded annotations. This balanced composition prevents bias and ensures robustness across various scenarios.

5. Conclusion

We present a flexible approach to amodal instance segmentation by adapting foundation segmentation models to handle both visible and occluded portions of objects. Our framework successfully leverages pre-trained modal detectors while maintaining strong amodal segmentation capabilities. The introduction of Amodal-LVIS, containing 300K carefully curated images, along with our comprehensive collection of 1M images and 2M instance annotations, addresses critical limitations in existing datasets and provides the necessary scale for robust model development.

Our extensive experiments demonstrate that SAMEO consistently outperforms state-of-the-art methods on COCOA-cls, D2SA, and MUVA datasets. Most notably, when trained on our dataset collection, including Amodal-LVIS, SAMEO achieves strong zero-shot performance on unseen datasets. The model's robust generalization abilities persist across various front-end detectors, validating our approach of adapting foundation models for amodal segmentation without compromising performance. We further address the limitations and possible future work of SAMEO in the appendix.

Acknowledgements

This work was supported in part by NSTC grants 113-2221-E-001-010-MY3 and 112-2221-E-A49-100-MY3 of Taiwan, as well as funding from NVIDIA Taiwan AI R&D Center. We appreciate the National Center for High-performance Computing for providing computational resources and facilities.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, pages 213–229. Springer, 2020. 2

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[3] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6144–6153. Computer Vision Foundation / IEEE Computer Society, 2018. 3, 4

[4] Patrick Follmann, Rebecca König, Philipp Härtinger, Michael Klostermann, and Tobias Böttger. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pages 1328–1336. IEEE, 2019. 2, 3, 4

[5] Jianxiong Gao, Xuelin Qian, Yikai Wang, Tianjun Xiao, Tong He, Zheng Zhang, and Yanwei Fu. Coarse-to-fine amodal segmentation with shape prior. In *ICCV*, pages 1262–1271. IEEE, 2023. 2

[6] Agrim Gupta, Piotr Dollár, and Ross B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5356–5364. Computer Vision Foundation / IEEE, 2019. 2

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988. IEEE Computer Society, 2017. 2

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE, 2022. 3

[9] Gaetano Kanizsa, Paolo Legrenzi, and Paolo Bozzi. Organization in vision: Essays on gestalt perception. 1979. 1

[10] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4019–4028. Computer Vision Foundation / IEEE, 2021. 2

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3992–4003. IEEE, 2023. 1, 2, 3

[12] Ke Li and Jitendra Malik. Amodal instance segmentation. In *ECCV*, pages 677–693. Springer, 2016. 2

[13] 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 *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 23447–23456. IEEE, 2023. 3, 4

[14] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3):3292–3310, 2023. 3

[15] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society, 2017. 3

[16] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. Rtmdet: An empirical study of designing real-time object detectors. *CoRR*, abs/2212.07784, 2022. 1, 2, 3, 6

[17] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*, pages 565–571. IEEE Computer Society, 2016. 3

[18] Rohit Mohan and Abhinav Valada. Amodal panoptic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20991–21000. IEEE, 2022. 3, 4

[19] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 3931–3940. IEEE, 2024. 3, 4

[20] Stephen E. Palmer. *Vision Science: Photons to Phenomenology*. MIT Press, 1999. 1

[21] Roberto Del Prete, Maria Daniela Graziano, and Alfredo Renga. Retinanet: A deep learning architecture to achieve

a robust wake detector in SAR images. In *6th IEEE International Forum on Research and Technology for Society and Industry, RTSI 2021, Naples, Italy, September 6-9, 2021*, pages 171–176. IEEE, 2021. 2, 6

[22] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with KINS dataset. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3014–3023. Computer Vision Foundation / IEEE, 2019. 2, 3, 4

[23] N. Dinesh Reddy, Robert Tamburo, and Srinivasa G. Narasimhan. WALT: watch and learn 2d amodal representation from time-lapse imagery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 9346–9356. IEEE, 2022. 3, 4

[24] Thomas F Shipley and Philip J Kellman. *From fragments to objects: Segmentation and grouping in vision*. Elsevier, 2001. 1

[25] Minh Q. Tran, Khoa Vo, Kashu Yamazaki, Arthur A. F. Fernandes, Michael 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*, page 712. BMVA Press, 2022. 2, 3, 6

[26] Haochen Wang, Xiaolong Jiang, Xu Tang, Yao Hu, Cilin Yan, Weidi Xie, Shuai Wang, and Efstratios Gavves. Towards open-vocabulary video instance segmentation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 4034–4043. IEEE, 2023. 2

[27] Liya Wang and Alex Tien. Aerial image object detection with vision transformer detector (vitdet). In *IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2023, Pasadena, CA, USA, July 16-21, 2023*, pages 6450–6453. IEEE, 2023. 2, 6

[28] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext V2: co-designing and scaling convnets with masked autoencoders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 16133–16142. IEEE, 2023. 1, 2, 6

[29] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest N. Iandola, Raghuraman Krishnamoorthi, and Vikas Chandra. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 16111–16121. IEEE, 2024. 1, 3

[30] Jianwei Yang, Zhile Ren, Mingze Xu, Xinlei Chen, David J. Crandall, Devi Parikh, and Dhruv Batra. Embodied amodal recognition: Learning to move to perceive objects. In *ICCV*, pages 2040–2050. IEEE, 2019. 2

[31] Jian Yao, Yuxin Hong, Chiyu Wang, Tianjun Xiao, Tong He, Francesco Locatello, David P. Wipf, Yanwei Fu, and Zheng Zhang. Self-supervised amodal video object segmentation. In *NeurIPS*, 2022. 2

[32] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. Amodal ground truth and completion in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 28003–28013. IEEE, 2024. 3, 4

[33] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2, 6

[34] Yan Zhu, Yuandong Tian, Dimitris N. Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3001–3009. IEEE Computer Society, 2017. 3, 4

[35] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 6725–6735. IEEE, 2023. 2, 6