MUVA: A New Large-Scale Benchmark for Multi-view Amodal Instance Segmentation in the Shopping Scenario

Zhixuan Li[1] Weining Ye[1] Juan Terven[2] Zachary Bennett[2]
Ying Zheng[2] Tingting Jiang[1,*] Tiejun Huang[1,3]
[1] National Engineering Research Center of Visual Technology, National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University, Beijing 100871, China
[2] AiFi Inc., California 94010, United States
[3] Beijing Academy of Artificial Intelligence, Beijing 100084, China
{zhixuanli,ywning}@pku.edu.cn, {juan,zachary}@aifi.com,
yingz@alumni.gsb.stanford.edu, {ttjiang,tjhuang}@pku.edu.cn

Abstract

*Amodal Instance Segmentation (AIS) endeavors to accurately deduce complete object shapes that are partially or fully occluded. However, the inherent ill-posed nature of single-view datasets poses challenges in determining occluded shapes. A multi-view framework may help alleviate this problem, as humans often adjust their perspective when encountering occluded objects. At present, this approach has not yet been explored by existing methods and datasets. To bridge this gap, we propose a new task called **M**ulti-view **A**modal **I**nstance **S**egmentation (MAIS) and introduce the MUVA dataset, the first **MU**lti-**V**iew **A**IS dataset that takes the shopping scenario as instantiation. MUVA provides comprehensive annotations, including multi-view amodal/visible segmentation masks, 3D models, and depth maps, making it the largest image-level AIS dataset in terms of both the number of images and instances. Additionally, we propose a new method for aggregating representative features across different instances and views, which demonstrates promising results in accurately predicting occluded objects from one viewpoint by leveraging information from other viewpoints. Besides, we also demonstrate that MUVA can benefit the AIS task in real-world scenarios.* [1]

1. Introduction

The amodal instance segmentation (AIS) task aims to determine an object's entire shape, encompassing its visible and occluded components. AIS task is more challenging than the visible instance segmentation task [23, 11] as it lacks occluded region appearance. Despite its complexity, the AIS task has significant implications for various industrial applications that encounter occlusion problems, such as robotic arm grasping [1], pedestrian re-identification [34, 36, 29], automatic driving [27, 28], and self-checkout systems in supermarkets [8].

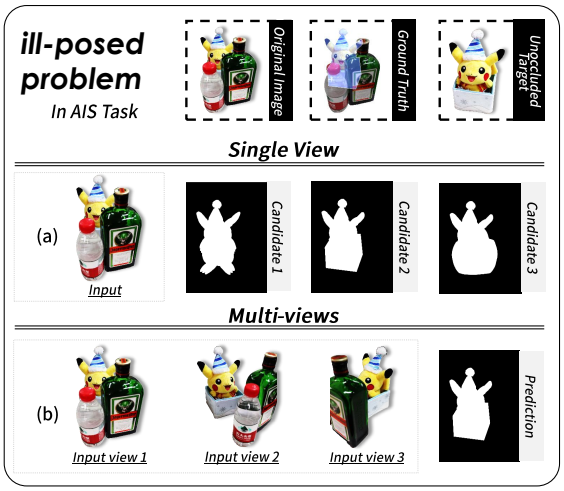

Figure 1. Comparison of the impact of ill-posed problems on amodal prediction in single-view and multi-view input settings. (a) In single-view input, ambiguity arises due to multiple candidates for the occluded object. (b) Multi-view input helps alleviate ambiguity and improves amodal prediction accuracy.

Although numerous datasets [19, 40, 8, 27, 13] and methods [19, 40, 8, 27, 13, 38, 33, 21] have been proposed since the AIS task was firstly introduced in 2016 by Li and Malik [19], current AIS datasets and methods rely on a *single-view* approach, suffering from the *ill-posed* problem. For example, as shown in Fig. 1(a), directly deducting

*Corresponding author
[1]The proposed MUVA dataset can be downloaded from this link: https://zhixuanli.github.io/project_2023_ICCV_MUVA.

the complete shape from a single-view image is extremely challenging due to the presence of occluded regions with multiple potential candidates. This is because distinct objects can share the same visible appearance but differ in shape within the occluded region. However, as shown in Fig. 1(b), humans tend to observe occluded objects from multiple angles to obtain accurate predictions.

Inspired by this observation, this paper proposes a novel task called **MAIS** (**M**ulti-view **A**modal **I**nstance **S**egmentation), which predicts amodal segmentations through multiple viewpoints. The task is designed for real-world applications, under the assumption that the number of viewpoints and field of vision is limited, and objects in the scene are closely distributed with reasonable occlusion. These assumptions contribute to the difficulty and practical significance of the MAIS task.

To study the MAIS task, creating a multi-view AIS dataset is essential. However, annotating multi-view data based on existing or newly collected real-world data involve significant efforts and may not be accurate in identifying shapes of occluded regions. This is because manually annotated amodal masks in real-world datasets are often inaccurate and inconsistent due to the varying shape prior knowledge of annotators, as shown in Fig. 2 (copied from [40]). Therefore, using a *synthetic* approach would be more precise and controllable. To create a synthetic dataset, one option is to build upon existing synthetic AIS datasets [6, 13, 14]. However, these datasets lack high-quality 3D models and sufficient occlusion. To overcome the limitation, we propose to create a new synthetic dataset comprising reconstructed high-quality 3D models and sufficient occlusion by controlling the distribution of objects. As an initial step in exploring the MAIS task, we limit our dataset construction to a single scenario. Specifically, the shopping scenario is selected due to its potential for multiple camera arrangements and its tendency for severe occlusion when goods are piled up. Consequently, we introduce **MUVA**, a novel **MU**lti-**V**iew **A**modal Instance Segmentation dataset. The dataset creation process follows a standard three-step approach. Initially, 3D artists construct models from images collected from on-sale items. Next, 3D models are then selected and placed in a 3D scene with careful compositions to control the occlusion degree. Finally, multi-view images are captured by six simulated cameras to simulate a real-world self-checkout setting. Each image is extensively annotated with visible/amodal segmentation masks, depth maps, occlusion orders, and 3D models.

To our best knowledge, MUVA is the first and only multi-view AIS dataset currently available. Unlike real-world-based AIS datasets [19, 40, 8, 27], MUVA provides precise annotations in the occluded region due to its synthetic nature and known ground-truth 3D models. Besides, MUVA is also the largest image-level AIS dataset in terms of both images and instances.

To exploit the multi-view information in MUVA, we introduce MASFormer, a novel method that aggregates information across different views and instances. Through comparative experiments with existing single-view-based AIS methods, we demonstrate that MASFormer significantly outperforms these methods by effectively utilizing multi-view information. Additionally, our approach can accurately predict objects that are severely occluded from one angle by incorporating information from other angles.

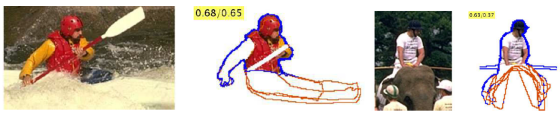

Figure 2. Two examples of COCOA [40] dataset, displaying multiple annotated occluded regions by different annotators.

Our contributions are summarized as follows: 1) **A new task named MAIS is proposed** to explore the AIS task under the multi-view setting for alleviating the ill-posed problem in the AIS task. 2) **A novel dataset named MUVA is proposed** for the MAIS task in the multi-view setting for shopping scenarios. To our best knowledge, MUVA is the first AIS dataset under the multi-view setting. 3) **A new method named MASFormer is proposed** to collect both view-level and instance-level information for solving the ill-posed problem in the AIS task. The experimental results show the efficiency of the multi-view setting over the single-view approach.

2. Related Work

2.1. Amodal Instance Segmentation Datasets

Various *single-view-based* AIS datasets have been proposed, including the image and video levels. (1) For the image level, most of the datasets are re-annotated manually based on the existing datasets for the visible instance segmentation (VIS) [5, 17, 30] task. For example, COCOA [40, 8], BSDSA [40], KINS [27] and D2SA [8] are extended based on the VIS datasets, including COCO [23], BSDS [25], KITTI [9] and D2S [7]. However, human annotation of occluded regions in image-level datasets may lack accuracy and consistency due to variations in the annotator's perception of amodal shape. Besides, DYCE [6] is a synthetic dataset for indoor furniture. (2) For the video level, two synthetic datasets, SAIL-VOS [13] and SAIL-VOS 3D [14] have been generated based on the GTA 3D game, containing daily-life scenes. Compared to video-level datasets with temporal consistency, MUVA exhibits higher shape variability across different viewpoints.

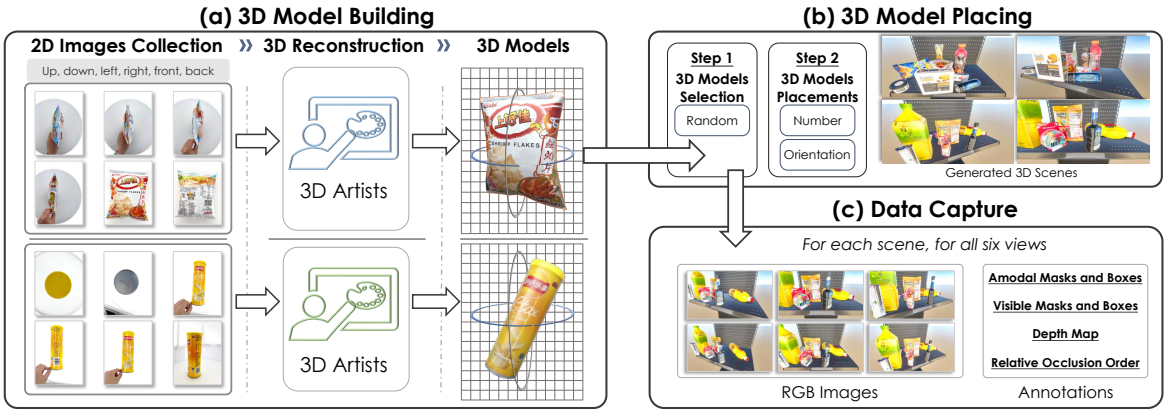

Figure 3. The pipeline of dataset generation. (a) For each object, 2D images are captured from up, down, left, right, front, and back, respectively. Then 3D artists use the collected images to reconstruct the 3D models. (b) For each scene, 3D models are randomly selected and placed with different amounts and orientations. (c) For each scene, six views are used to capture the data, including the RGB images and various annotations.

2.2. Amodal Instance Segmentation Methods

AIS methods employ various cues for amodal completion and can be categorized into three types based on relevant literature. First, numerous methods are adapted from VIS methods. Li and Malik [19] use Iterative Instance Segmentation [18] to solve the amodal problem. ORCNN [8] extends Mask-RCNN for simultaneous visible and amodal segmentations. ASN [27] learns features with a proposed multi-level aggregation module. OAFormer [20] proposes to learn the occlusion discriminative queries with a transformer-base network. Second, some studies utilize the relative occlusion order between objects, including SLN [38] predicts and combines the relative orders, and Deocclusion [35] employs a self-training method to learn an ordering map for guiding the amodal completion. Third, some studies use the shape prior knowledge to help infer the shape of the occluded region, such as Amodal-VAE [24], ShapeDict [33], GIN [22] and A3D [21]. However, the multi-view information is not utilized in existing methods due to the absence of a suitable dataset. Based on the proposed MUVA dataset, this paper presents a 2D approach to leverage multi-view knowledge. Instead of using sophisticated 3D approaches like reconstruction, our 2D approach can be easily applicable to real-world scenarios without requiring 3D supervision signals.

3. MUVA: Multi-view Amodal Dataset

MUVA comprises 1,801 distinct 3D objects reconstructed from multi-view photos. These objects are utilized to fabricate 4,401 scenes with diverse object placements. 26,406 RGB images are then produced from six cameras, each from differing viewpoints. The dataset includes thorough annotations, such as depth maps, 3D models, and amodal/visible masks. The dataset split of MUVA, including train, validation and test, are shown in the supplementary material.

3.1. Dataset Contents

The contents of MUVA including images and corresponding annotations are introduced below. RGB images are captured from each camera view with a resolution of 1920×1080. The high-quality 3D reconstructed models lead to final rendered RGB images with rich scene details and high visual quality. Six types of annotations are available for each RGB image. (1) The segmentation annotations of instances contain both *visible* and *amodal* instance masks. These are generated by the 3D simulation software, resulting in higher accuracy compared to manual annotations. Moreover, the occluded region mask annotations are more precise compared to human-annotated amodal masks which may vary between annotators. (2) *Depth maps* are produced to reflect the value of each pixel with regard to its distance from the relevant camera to the 3D object. The depth map registers distance for visible regions exclusively, excluding amodal regions. (3) The *occlusion order* specifies the list of other objects occluded by this one. (4) Accompanying high-quality 3D models are included. (5) The *bounding boxes* and *area* of both visible and amodal masks derived from corresponding masks are provided.

3.2. Dataset Generation

A universal pipeline is designed to generate the synthetic dataset MUVA in three stages: building 3D models, placing 3D models, and capturing data, as illustrated in Fig. 3.

3D Model Building. In this stage, 3D models are created as the foundation for the 3D scene, as presented in Fig. 3 (a). This stage contains two steps: (1) capturing six differ-

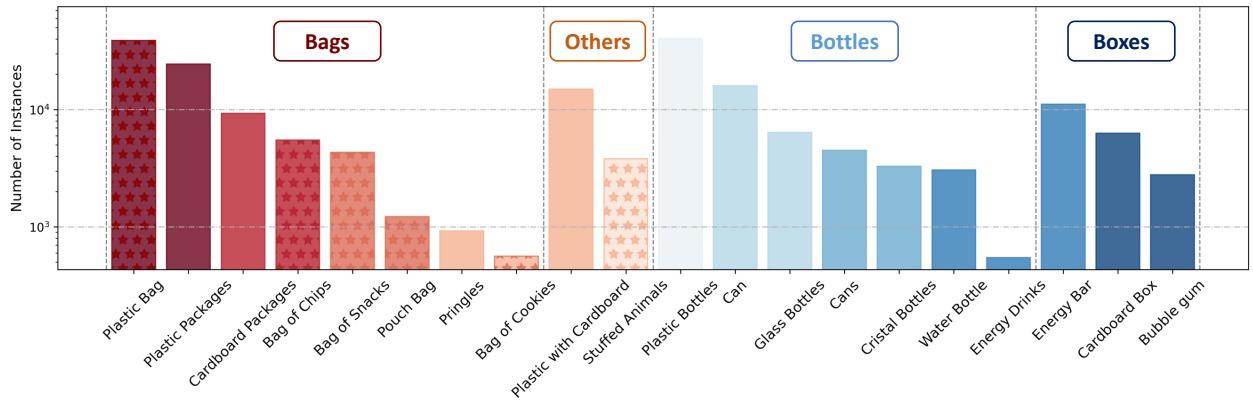

Figure 4. The histogram of the number of instances for four parent categories (bags, others, bottles, boxes) and corresponding sub-categories. Non-rigid categories are noted with star stripes. Best viewed in color.

Dataset	Publication	Data Type	Multiple Views	# Image	# Instance	Scenario	Synthetic or Real	3D Model	Depth	# Categories	Resolution (pixels)
SAIL-VOS [13]	CVPR'19	video	-	111,654	1,896,295	daily life	synthetic	-	✓	162	1M
SAIL-VOS 3D [14]	CVPR'21	video	-	237,611	3,460,213	daily life	synthetic	✓	✓	178	1M
COCOA [40]	CVPR'17	image	-	5,000	46,314	daily life	real	-	-	N/A	0.28M
BSDSA [40]	CVPR'17	image	-	500	3,739	daily life	real	-	-	N/A	0.15M
DYCE [6]	CVPR'18	image	-	5,500	85,975	daily life	synthetic	-	-	79	1M
KINS [27]	CVPR'19	image	-	14,991	190,626	street	real	-	-	8	0.47M
D2SA [8]	WACV'19	image	-	5,600	28,720	shopping	real	-	-	60	**3M**
COCOA-cls [8]	WACV'19	image	-	3,501	10,562	daily life	real	-	-	**80**	0.28M
MUVA	N/A	image	✓	**26,406**	**198,573**	shopping	synthetic	✓	✓	20	2M

Table 1. Comparison with existing amodal instance segmentation datasets. # means the number of this item. Bold numbers denote the largest one in each column among image-level datasets.

ent images of each object to provide adequate information for 3D reconstruction, and (2) utilizing MAYA, a 3D modeling software, to reconstruct and manually refine the 3D objects iteratively by 3D artists. This process results in the reconstruction of 1801 distinct 3D objects, each with reconstructed shape and texture.

3D Model Placing. This stage involves generating 3D scenes by selecting and placing 3D object models. It comprises two steps: 3D model selection and placement, as depicted in Fig. 3 (b). First, a shelf is created as a platform for scene construction, and 3D models are randomly chosen from those reconstructed in the previous stage. Second, placement settings, including object number, orientation, and position, are varied randomly to enhance diversity. The number of objects can range from less than 10 to more than 10, and their upright or flat placement is randomized. Object positions are also randomized, ensuring they are within the shelf's range. Finally, Unity3D software is used to render the scene. The physical simulation utilizes non-deformable 3D models with collision detection to prevent object penetration. The scenes were illuminated uniformly and gravity was applied to mimic realistic conditions. Creating all 1801 3D models required approximately 600 hours with a 20-minute time frame for each model. The cost of generating the models totaled approximately 15,000 USD at a rate of 25 USD per hour for each 3D artist.

Data Capture. The simulated scene depicted in Fig. 3 (c) provides various types of data for sampling. Six cameras are positioned in the front to capture sparse views and simulate the self-checkout setting, ensuring the applicability of MUVA to real-life scenarios. For each view, RGB images and numerous annotations are generated.

The above dataset generation pipeline is adaptable for creating datasets for other vision tasks, such as indoor 3D reconstruction [4] and person re-identification [31].

3.3. Dataset Statistics

MUVA dataset comprises 26,406 images with 198,573 instances categorized into four parent categories (bags, bottles, boxes, and others), and 20 sub-categories that include both rigid and non-rigid objects. The instance ratio of each category is presented in Fig. 4. The dataset contains 27.41% non-rigid objects, such as the "Plastic Bag," which provides shape variations. The proposed dataset includes high-resolution images of 1920×1080 pixels, providing rich details of the objects' edges and surfaces.

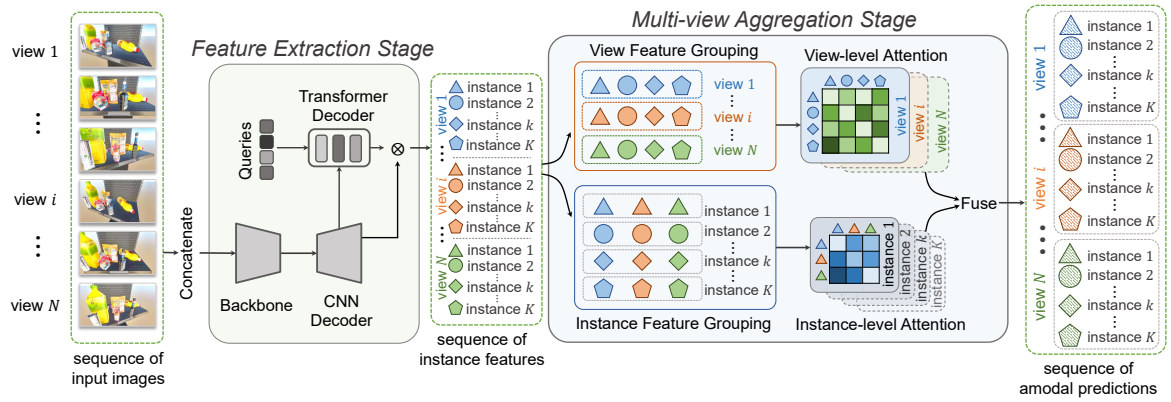

Figure 5. Overall architecture of our method. There are two stages in our method, including the Feature Extraction Stage and the Multi-view Aggregation Stage. The first stage aims to extract the features of each instance from all views. The second stage utilizes the relationship of features at the instance level and view level, respectively, to generate accurate predictions.

3.4. Comparison with Existing Datasets

Tab. 1 presents a comparison between MUVA and other AIS datasets, including video and image-level datasets such as COCOA [40], BSDSA [40], DYCE [6], KINS [27], D2SA [8] and COCOA-cls [8]. MUVA stands out as the only dataset specifically designed for the multi-view AIS task and the only synthetic image-level dataset that provides 3D models. The 3D models used to build the dataset are publicly available to support new methods for the AIS task. Additionally, MUVA and D2SA are the only two datasets for the shopping scenario. MUVA's synthetic nature leads to consistent and higher-quality annotations than human-annotated datasets. Moreover, MUVA has the largest number of images and instances among image-level datasets, and is the only dataset at the image level that contains depth maps, which can potentially help with the AIS task.

Dataset	simplicity ↓		convexity ↓	
	visible	amodal	visible	amodal
BSDS-A	0.718	0.834	0.616	0.643
COCO-A	0.746	0.856	0.658	0.685
KINS	0.709	0.830	0.610	0.639
MUVA	0.959	0.801	0.923	0.963

Table 2. Complexity is evaluated through simplicity and convexity metrics. Small values of these metrics indicate a complex dataset. A circle shape yields the maximum value of 1.0 for both metrics.

Shape Complexity. To assess the shape complexity of mask annotations contained in MUVA, metrics proposed by COCOA [40] is used. These metrics include convexity and simplicity for both visible and amodal masks. As shown in Tab. 2, MUVA has the lowest shape simplicity (0.801) and the highest convexity (0.963) of amodal masks. The high convexity can be attributed to the presence of goods such as bags, bottles, and boxes, which possess naturally convex shapes. Conversely, the low simplicity confirms that the shapes within MUVA are complex. For instance, a plastic bag with zigzagged boundaries exhibits high convexity but low simplicity.

3.5. Applications

MUVA provides diverse annotations that can be used for multiple research tasks. For example, using instance IDs from different viewpoints can lead to a new task of multi-view amodal instance segmentation (MAIS), which involves matching and segmenting the same instance from multiple viewpoints. High-quality 3D models can be used for single/multiple-view 3D reconstruction. The provided relative depth order annotation allows for occlusion order prediction. Finally, the depth map from the object to the camera can be utilized for depth estimation.

4. Method

This paper focuses on the newly formulated multi-view amodal instance segmentation (MAIS) task. This section proposes a network designed for the MAIS task, named Multi-view Amodal Segmentation Transformer (MASFormer). The task definition, overall architecture, and the proposed new modules are introduced in the following.

4.1. Overview

Problem Definition. Given N input images from N viewpoints, the MAIS task aims to predict K instance-level segmentation masks for interested objects in each image.

Overall Architecture. The whole pipeline of the proposed method contains two stages, including the Feature Extraction Stage (FES) and the Multi-view Aggregation Stage (MAS). (1) The FES extracts features of each instance

from all angles. (2) The MAS takes the instances' feature from FES as input and optimizes from two aspects, including aggregating all instances' features of the same view and all views' features of the same instance. Finally, the amodal instance predictions of objects in all images are generated after fusion.

4.2. Feature Extraction Stage

This stage aims to extract features from all views simultaneously. As shown in Fig. 5, the *Backbone* takes for N images as input and extracts representative features for K instances, including low-level structural details and high-level semantic information. Next, the *CNN Decoder* up-samples the representative features and generates multi-level features containing all objects' information. Then a *Transformer Decoder* is employed to output features F_{FES} corresponding to *all objects in all views*. Bipartite matching [2] is used to find the correspondence between predictions and supervisions, to specify the *view id* and *instance id* of each feature.

4.3. Multi-view Aggregation Stage

As shown in Fig. 5, this stage aims to enhance the extracted features F_{FES} from two aspects, including view-level and instance-level. Two branches are designed for the two aspects, respectively.

View-level Aggregation. This branch aims to utilize the relationship between features of all instances in the same view for enhancing each instance's feature quality. Firstly, for input features F_{FES} including *all objects in all views*, an *View Feature Grouping* module takes features of the same view into the same group. Then the *View-level Attention* module learns the correlation between all instances' features in the same view. Finally, the feature of each instance is enhanced by integrating all correlated instances' features in the same view. Here the correlation between different features is computed by cosine similarity. And the self-attention [32] is used to construct this module. Please refer to our supplementary material for detailed explanations of the network. Besides, this module requires objects in different views could be identified and matched. To achieve this, for each instance, both the view id and the instance id are predicted by the network and supervised by using the bipartite matching [2] to match between the prediction and ground truth.

Instance-level Aggregation. This module aims to use features of the same instance in all views to improve the feature quality of each instance. First, the *Instance Feature Grouping* module takes features of the same instance across all views to the same group. Next, for each group, the *Instance-level Attention* module learns correlations between features of the same instance in different views and integrates associated features by using the self-attention [32] mechanism.

Finally, a CNN-based fuse layer is used to merge the output features of two branches for the final prediction of amodal masks. With the aggregation of two branches, the feature representations of all instances can be enhanced, and more precise amodal segmentation results can be obtained.

5. Experiments

This section first compares the newly proposed method MASFormer with existing single-view-based state-of-the-art AIS methods on the proposed MUVA dataset. Then the ablation studies are conducted to show the effectiveness of amount and order for the input viewpoints and components of the proposed method.

Method	Publication	AP	AP_{50}	AP_{75}	AR
Mask-RCNN [11]	ICCV'17	17.2	26.6	18.7	<u>48.8</u>
ORCNN [8]	WACV'19	7.8	25.9	2.1	24.7
SLN [38]	ACM MM'19	6.9	20.2	5.6	22.9
DeepSnake [26]	CVPR'20	7.7	11.9	7.4	23.8
BCNet [16]	CVPR'21	23.1	28.9	25.0	40.9
ShapeDict [33]	AAAI'21	9.9	36.0	1.3	27.8
E2EC [37]	CVPR'22	11.8	15.7	12.9	25.5
Baseline (1 view)	CVPR'22	17.8	26.1	18.0	29.2
Baseline (6 views)	CVPR'22	22.9	35.5	22.3	36.4
Ours (1 view)	N/A	<u>25.6</u>	<u>38.0</u>	<u>25.6</u>	41.2
Ours (6 views)	N/A	**39.4**	**55.9**	**40.3**	**51.8**

Table 3. Compared with state-of-the-art methods on MUVA. In each column, the best is in **bold**, and the second-best is <u>underlined</u>.

Method	ResNet-50	ResNet-101	ResNeXt-101
MaskRCNN	17.2	19.1	19.8
BCNet	23.1	25.3	26.6
E2EC	11.8	14.2	15.9
Ours (6 views)	39.4	42.5	43.0

Table 4. Performance of methods with stronger backbones.

Experimental Settings. The baseline method only contains the first stage FES in the proposed method MASFormer, which is implemented following the video instance segmentation method Mask2Former [3] as an instantiation. The baseline method can take multi-view images as input and outputs amodal segmentation masks, the same as MASFormer. ResNet-FPN-50 [12] is employed as the backbone for all methods. The results are evaluated using the AP and AR metric, commonly used to measure the precision and recall of prediction. For fairness, all experiments use the same training dataset of MUVA to train and the same validation dataset of MUVA to validate. The ground truth of all methods is the same, including the bounding boxes and masks for amodal and visible regions, image ids, and instance ids. The ground-truth camera viewpoint information and 3D models are *not* used for fairness.

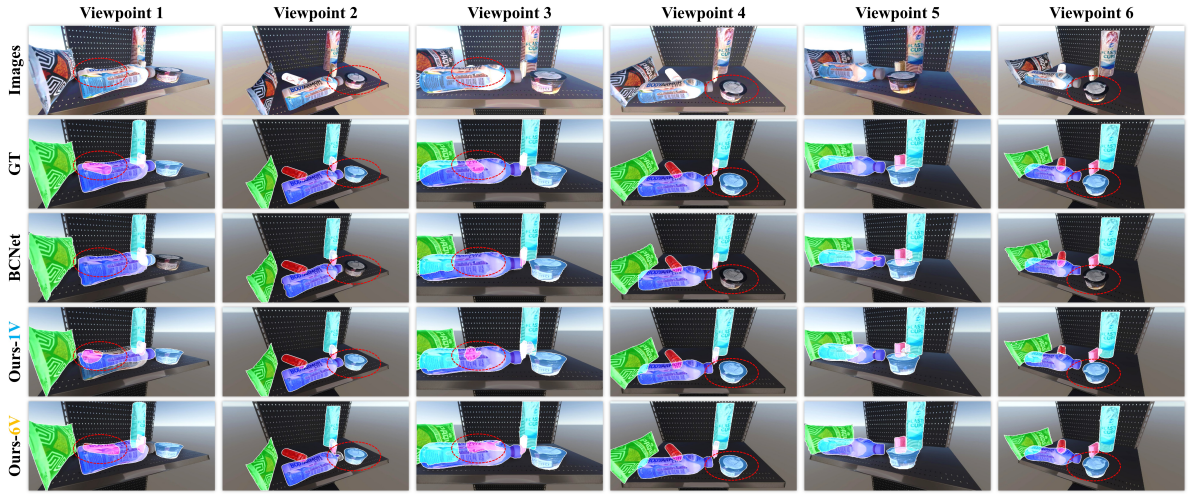

Figure 6. Visualization comparisons between BCNet [16] and ours on MUVA, trained with one viewpoint (1V) and six viewpoints (6V). For masks, different colors denote different instances, and the same instance in different angles has the same color. Red circles indicate regions should be focused. Zoom in for a better view. In the first and third columns, even if a bottle is severely occluded, it can be predicted accurately by our method. Moreover, our method trained with 6 views performs better than training with a single view.

5.1. Multi-view Amodal Instance Segmentation

Tab. 3 shows the comparison results of some existing AIS methods on MUVA. Several state-of-the-art methods designed for single-view datasets are compared. For baseline and the proposed method MASFormer, two kinds of input are used during training, one is a single view, and the other is six. Note that the amount of training data is the same for these two settings.

We compare the results at three levels, including *one-view*, *one vs. six views*, and *six-view* levels. (1) As shown in Tab. 3, when only using *one view* for training, our proposed method MASFormer achieves the best performance (25.6% AP, 38.0% AP_{50}, 25.6% AP_{75}, 41.2% AR) on all metrics. MASFormer outperforms the best performance of *single-view-based* methods with 2.7% AP, 2.0% AP_{50}, 0.6% AP_{75}, and 0.2% AR. The results demonstrate the effectiveness of the proposed *view-level* attention module, which can learn the relationship between different instances of the same view and improve the amodal prediction with the help of a learned relationship. (2) Both ours and the baseline method achieve better performance when using six views than one view for training. The results show that even with the same training data, using more views as input for training can improve the prediction. (3) When both methods use *six views* for training, our method can outperform the baseline method with 16.5% AP. The results show the effectiveness of the proposed *instance-level* attention module, which learns and benefits from the relationship between different views of the same instance. Besides, the performance of methods with stronger backbones is shown in Tab. 4. Stronger backbones bring improvements for all methods, while the proposed MASFormer still achieves the best performance. Qualitative results are shown in Fig. 6.

Setting	Training Dataset	Testing Dataset	AP
1	$\text{D2SA}_{\text{train}}$	D2SA_{val}	63.5
2	$\text{MUVA}_{\text{train}}$	D2SA_{val}	41.9
3	$\text{D2SA}_{\text{train}}$ + $\text{MUVA}_{\text{train}}$	D2SA_{val}	68.4

Table 5. Performance of MASFormer evaluated on the validation set of D2SA dataset and trained with different combinations of source-domain datasets for training. For the two datasets, only data of overlapped categories are used.

5.2. Generalization to Real-world Scenarios

In this section, we demonstrate the generalization ability of the proposed dataset and method on real-world scenarios under single-view and multi-view settings, respectively.

Single-view Generalization. To validate generalization ability in real-world scenarios, the MUVA dataset is utilized under a single-view setting. To be specific, we employ settings of the domain generalization [39, 10, 15] task to evaluate the generalization ability of learning from a source domain, typically synthetic datasets, to out-of-distribution target domains like real-world datasets. Specifically, the MUVA synthetic dataset and real-world AIS dataset D2SA [8] are used in this study. D2SA contains 60 categories, of which 18 categories overlap with MUVA. More information about the category labels can be found in the supplementary material. The following experiments only consider data from the overlapping categories. Three settings were adopted as presented in Tab. 5. Firstly, MASFormer is trained and evaluated on the train and validation sets of

D2SA datasets, demonstrating its ability to generalize well on real-world data. Secondly, MASFormer is trained on the synthetic MUVA dataset and evaluated on the real-world dataset D2SA, with results validating the generalizability of the synthetic MUVA dataset. Lastly, mixing D2SA and MUVA for training lead to a 4.9% AP improvement, indicating the benefits of using MUVA in a real-world scenario.

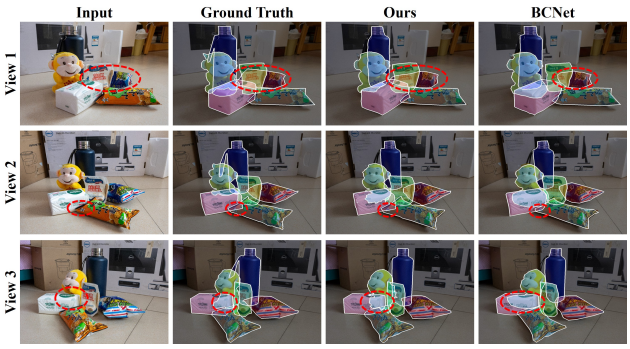

Figure 7. Qualitative results of a real-world case. Red circles indicate differences between ours and BCNet. Zoom in for a better view.

Multi-view Generalization. Considering there is no existing multi-view real-world AIS dataset, we construct a *real-world* shopping scenario to obtain multi-view data for validating the generalization ability of the proposed dataset and method. Images are captured from the left, middle, and right sides, respectively. MASFormer trained with six views and BCNet are used. And the train set of MUVA is utilized for training. As shown in Fig. 7, MASFormer can obtain good amodal segmentation results, demonstrating the generalization ability of MUVA to real-world scenarios. Moreover, when compared with BCNet, MASFormer performs better consistently on all viewpoints, validating its potential for generalizing to the real-world scene.

Views Order	# Views	AP	AP_{50}	AP_{75}	AR
N/A	1	25.6	38.0	25.6	39.2
Fixed Order	2	26.2	38.0	26.3	39.8
	3	28.0	42.9	27.7	40.9
	4	30.9	45.6	30.9	43.5
	5	34.1	49.0	34.7	47.4
	6	<u>37.9</u>	<u>53.9</u>	<u>38.6</u>	<u>51.4</u>
Random Order	2	26.9	39.5	27.1	40.4
	3	28.4	41.7	28.4	42.9
	4	32.8	46.6	33.4	45.6
	5	36.9	52.7	37.2	50.5
	6	**39.4**	**55.9**	**40.3**	**51.8**

Table 6. Effect of fixing views input order and views amount. In each column, the best is in **bold**, and the second-best is <u>underlined</u>.

5.3. Ablation Study

We conducted ablation study experiments for MASFormer, examining the impact of the number and order of input viewpoints, the effectiveness of attention modules, and computational cost.

Viewpoints Count and Order. Tab 6 presents some combinations of the number and order of input viewpoints used for training in the proposed MASFormer method. The same amount of training data is used for all experiments. Performance consistently improves as the number of input views increases, regardless of whether their order is fixed (rows # 2 to # 6) or random (rows # 7 to # 11). These results show that using more views for training can improve performance and multi-view data is crucial for amodal instance segmentation. When the number of input views is constant, the impact of the input views' order is analyzed below. As shown in Tab. 6, results of random input order outperform fixed order ones with the equal number of input views. For example, random order improves the AP metric with the following percentages compared to fixed order: 0.7% (2 views), 0.4% (3 views), 1.9% (4 views), 2.8% (5 views) and 1.5% (6 views). Results show that randomizing input image ordering of various viewpoints leads to increased training sample distribution and improved effectiveness.

Index	View-level	Instance-level	AP	AP_{50}	AP_{75}	AR
1			22.9	35.5	22.3	36.4
2	✓		28.7	43.1	28.3	41.6
3		✓	<u>34.1</u>	<u>49.0</u>	<u>34.7</u>	<u>47.4</u>
4	✓	✓	**39.4**	**55.9**	**40.3**	**51.8**

Table 7. Effect of two attention modules in the proposed method.

Attention Modules. Tab. 7 shows the ablation study of proposed attention modules proposed in our MASFormer method, using six input views with random order. The proposed view-level attention module improves AP performance from 22.9% to 28.7% by learning feature relationships within the same view. Instance-level attention module achieves an 11.2% improvement by utilizing multi-view features of the same instance. Both attention modules combined result in a 39.4% AP performance, surpassing individual settings.

Method	Framework	GPU Mem	FPS	Time	AP
BCNet	CNN	4G	8.3	4h	23.1
Baseline	Transformer	9G	6.5	6h	22.9
Ours	Transformer	13G	3.7	8h	39.4

Table 8. Comparison of computational cost and the performance. Time means the training time.

5.4. Computational Cost

Tab. 8 compares the computational cost. The ResNet50-FPN backbone is used for all methods. The FPS is the average of all images in the validation dataset. Compared to other methods, our method exhibits superior performance at

the expense of higher computational cost due to the use of Transformer framework.

6. Limitation and Conclusion

Limitations of this paper include the consideration of only one scenario, shopping, and a high computational cost of the proposed method. Future research should aim to expand the dataset to include other scenarios and develop a more efficient network that incorporates additional information such as 3D models from MUVA. As for the conclusion, this paper introduces a novel task called multi-view amodal instance segmentation (MAIS) in the shopping scenario. To facilitate research in this area, a new synthetic dataset, MUVA, is introduced along with a proposed method that utilizes multi-view information.

Acknowledgments

This work was partially supported by the Natural Science Foundation of China under contract 62088102. This work was also partially supported by Qualcomm. We also acknowledge High-Performance Computing Platform of Peking University for providing computational resources.

References

[1] 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 *International Conference on Robotics and Automation*, pages 5085–5092, 2022.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 213–229, 2020.

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017.

[5] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. CenterNet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6569–6578, 2019.

[6] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. SeGAN: Segmenting and generating the invisible. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6144–6153, 2018.

[7] Patrick Follmann, Tobias Bottger, Philipp Hartinger, Rebecca Konig, and Markus Ulrich. MVTec D2S: Densely segmented supermarket dataset. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 569–585, 2018.

[8] Patrick Follmann, Rebecca König, Philipp Härtinger, Michael Klostermann, and Tobias Böttger. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1328–1336, 2019.

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[10] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2961–2969, 2017.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[13] Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, and Alexander G Schwing. SAIL-VOS: Semantic amodal instance level video object segmentation - a synthetic dataset and baselines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3105–3115, 2019.

[14] Yuan-Ting Hu, Jiahong Wang, Raymond A. Yeh, and Alexander G. Schwing. SAIL-VOS 3D: A synthetic dataset and baselines for object detection and 3D mesh reconstruction from video data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 3359–3369, 2021.

[15] Xueying Jiang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Domain generalization via balancing training difficulty and model capability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[16] 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*, pages 4019–4028, 2021.

[17] Youngwan Lee and Jongyoul Park. CenterMask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13906–13915, 2020.

[18] Ke Li, Bharath Hariharan, and Jitendra Malik. Iterative instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3659–3667, 2016.

[19] Ke Li and Jitendra Malik. Amodal instance segmentation. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 677–693, 2016.

[20] Zhixuan Li, Ruohua Shi, Tiejun Huang, and Tingting Jiang. OAFormer: Learning occlusion distinguishable feature for amodal instance segmentation. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1–5, 2023.

[21] Zhixuan Li, Weining Ye, Tingting Jiang, and Tiejun Huang. 2D amodal instance segmentation guided by 3D shape prior. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 165–181, 2022.

[22] Zhixuan Li, Weining Ye, Tingting Jiang, and Tiejun Huang. GIN: Generative invariant shape prior for amodal instance segmentation. In *IEEE Transactions on Multimedia*, 2023.

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 740–755, 2014.

[24] Huan Ling, David Acuna, Karsten Kreis, Seung Wook Kim, and Sanja Fidler. Variational amodal object completion. In *Advances in Neural Information Processing Systems*, pages 16246–16257, 2020.

[25] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 416–423, 2001.

[26] Sida Peng, Wen Jiang, Huaijin Pi, Xiuli Li, Hujun Bao, and Xiaowei Zhou. Deep snake for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[27] 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.

[28] Mennatullah Siam, Sara Elkerdawy, Martin Jagersand, and Senthil Yogamani. Deep semantic segmentation for automated driving: Taxonomy, roadmap and challenges. In *IEEE International Conference on Intelligent Transportation Systems*, pages 1–8, 2017.

[29] Xiaolin Song, Kaili Zhao, Wen-Sheng Chu, Honggang Zhang, and Jun Guo. Progressive refinement network for occluded pedestrian detection. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, 2020.

[30] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 282–298, 2020.

[31] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4627–4635, 2017.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[33] Yuting Xiao, Yanyu Xu, Ziming Zhong, Weixin Luo, Jiawei Li, and Shenghua Gao. Amodal segmentation based on visible region segmentation and shape prior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2995–3003, 2021.

[34] Jin Xie, Yanwei Pang, M. H. Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Mask-guided attention network and occlusion-sensitive hard example mining for occluded pedestrian detection. *IEEE Transactions on Image Processing*, 30:3872–3884, 2021.

[35] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene deocclusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3784–3792, 2020.

[36] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and S. Li. Occlusion-aware R-CNN: Detecting pedestrians in a crowd. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, 2018.

[37] Tao Zhang, Shiqing Wei, and Shunping Ji. E2EC: An end-to-end contour-based method for high-quality high-speed instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4443–4452, 2022.

[38] Ziheng Zhang, Anpei Chen, Ling Xie, Jingyi Yu, and Shenghua Gao. Learning semantics-aware distance map with semantics layering network for amodal instance segmentation. In *Proceedings of the ACM International Conference on Multimedia*, pages 2124–2132, 2019.

[39] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021.

[40] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1464–1472, 2017.