Supplementary Material of "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

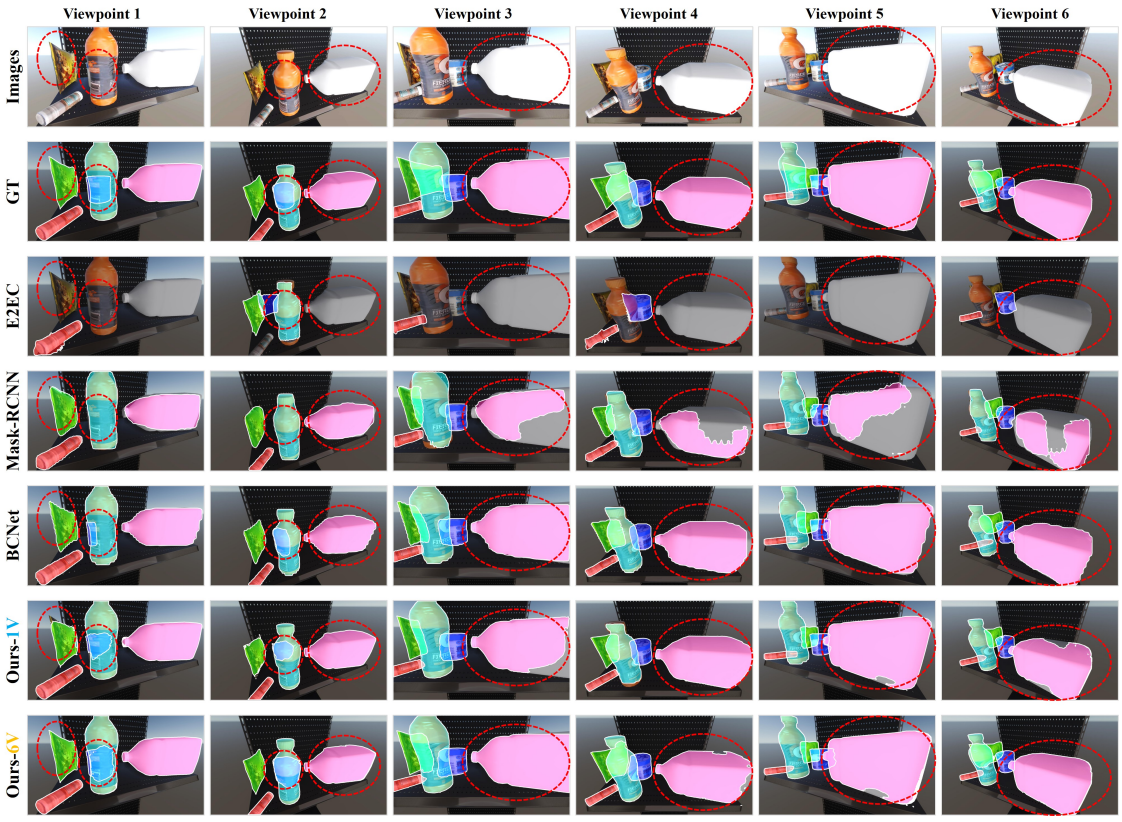

Figure 1: The first example of comparison between state-of-the-art methods and ours on MUVA, trained with one viewpoint (1V) and six viewpoints (6V). For masks, different colors denote different instances, and the same instance in different angles has the same color. Red circles indicate regions should be focused. Zoom in for a better view.

This supplementary material provides additional qualitative results, discussions and additional details, including the proposed dataset, method.

* Corresponding author

1. Qualitative Results

In this section, results of *two* example scenes are given to make a qualitative comparison between the proposed method and the single-view-based methods, including **E2EC**, **Mask-RCNN**, and **BCNet**.

The visualizations and corresponding analysis of the comparisons are shown in the following.

Fig. 1 shows an example of comparison between single-view and our proposed multi-view AIS methods. Different rows show the images, ground-truth (GT) amodal masks, and predicted amodal masks of different methods. Each row has six views, and each column contains the same view. Note that all experiments use the same data for training and testing, making the comparison fair among all methods. As shown in the first row, two objects need to be focused on. The first one is the plastic bag on the left (first row, first column, pointed by the red arrow), and the second one is the ice cream occluded by the orange juice bottle (first row, third column, pointed by the red arrow). In the following, there are some observations of the comparison.

The first object, the plastic bag, is heavily occluded in the last column. Moreover, our proposed multi-view method can predict a better amodal mask for the occluded chip bag than all other methods. The first column shows a comparison of the segmentation results for all methods for the ice cream bottle. Our method with six views achieves the best performance for the ice cream bottle, while E2EC and Mask-RCNN do not find the occluded ice cream bottle. BCNet can find the bottle but failed to predict the accurate amodal mask. The results show that multi-view methods can merge the knowledge from other viewpoints for helping to predict the complete shape of the occluded viewpoints. Our method trained with six views also outperforms all other methods.

Fig. 2 shows another example of a comparison between single-view and our proposed multi-view AIS methods. The first row shows the images of all six views. The second row shows all six views' ground-truth (GT) amodal masks. The rest rows show the prediction of different methods, including all views. Three objects need to be focused on. As shown in the first row, the first object is the plastic bag (first column, pointed by a red arrow). The second and third objects are two bottles, shown in the sixth column of the first row (pointed by a red arrow). In the following, there are some observations of the comparison.

The first object, the plastic bag, is heavily occluded by the coffee bottle, as shown in the last column. The comparison of results from different methods shows that our proposed multi-view method can achieve the best performance. Our method can take advantage of the multi-view images and complement each other for the occluded parts.

For the second and third objects, the two bottles are hard to see in the first column. The E2EC and Mask-RCNN methods predict only one bottle's amodal mask. Moreover, the BCNet method does not find any occluded bottles behind the coffee bottle. However, our method can find both occluded bottles and predicts accurate amodal masks, especially when using six views for training (as shown in the last row). Similar performance appears in the second, third, and fourth columns.

2. Discussions

2.1. Difficulty of the proposed MAIS task

The MAIS (Multi-view Amodal Instance Segmentation) task could be easy under ideal conditions if there are enough views and the objects are sparsely located. However, it is not easy to satisfy both preconditions in real life, including enough views and sparse locations. We propose two assumptions to ensure the difficulty of the MAIS task, including limited observation of the scene and reasonable occlusion between objects in the scene. Moreover, both ideal conditions do not hold in our proposed dataset MUVA.

2.2. Technical difference between video and multi-view AIS datasets

The main difference between video and multi-view tasks is that the assumptions of the consistency for the two tasks are different. For the video task, different frames are temporally consistent. However, for the multi-view task, different images for the same scene are consistent in multiple views.

Specifically, for temporal consistency, the shapes and positions of the same object in the 3D space (like a human) across different frames may be different. However, in the MAIS task, all views are captured simultaneously, which means the shapes and positions of each object in the 3D space across all views are the same. Therefore the proposed multi-view AIS dataset is different from video-level AIS datasets.

2.3. Applicability of the dataset construction method for other scenarios

The dataset construction method is a typical pipeline, including 2D data collection, 3D model reconstruction, 3D model placements, and data capturing. This pipeline could be applied for many scenarios, such as Crowd Counting [1] and Person Re-ID [2].

2.4. Reason of building a synthetic dataset

The amodal annotations of a real-world dataset are challenging to obtain and can not be accurate, no matter whether it is *newly constructed* or *extended* based on existing datasets. For example, for a street scene with persons and cars, obtaining the ground-truth amodal mask of an occluded person is challenging. This is because the appearance of the person's occluded region is unavailable, and the

Figure 2: The second example of comparison between state-of-the-art methods 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.

shape of the occluded region is hard to determine by the annotator.

2.5. Reason of collecting data from one scenario

MUVA dataset contains data from only one scenario at present because this dataset serves as *the first step* to explore the multi-view amodal segmentation task. In the future, new datasets collected from other scenarios could be considered to extend the applications.

2.6. Reason of selecting the shopping scenario

The shopping scenario is chosen for several reasons.

First, the self-service supermarket is trendy. Amazon and Walmart have been promoting the self-service supermarket for several years.

Second, there are many techniques required to supply the self-servicing. One of the most critical requirements is auto-checking, which needs to obtain the final prices of all piled objects that a customer has picked. The sensor-based method (like RFID label or weight sensor) is expensive, and the vision-based method (only using the camera) is cheap. Moreover, for the vision-based method, one of the biggest challenges is the occlusion problem caused by stacking all of the purchases. The amodal instance segmentation task is a helpful method to handle this problem.

Thirdly, obtaining data from multiple cameras in shopping scenarios is convenient.

Finally, the previously commonly used dataset D2SA also concentrates on the shopping scenario.

Therefore, we believe the shopping scenario is an important application. For example, from the aspect of tasks, self-checking and security are essential applications. From the aspect of techniques, 3D reconstruction and depth estimation for the shopping scenarios are both meaningful.

3. Dataset

3.1. Dataset Split

The entire MUVA dataset has 4401 scenes, split into three parts: train, validation, and test. Tab. 1 shows the detailed numbers of images and instances in each part. The ratio of the scenes is 3:1:1 for train, validation, and test, respectively.

Table 1: The number of images and instances in each dataset split. # means the number of this item.

	# Image	# Instance	# Scene
Train	16008	119712	2640
Validation	5166	39390	880
Test	5232	39471	881

4. Method

This section provides details for the proposed method, including the grouping and attention modules in the Multi-view Aggregation Stage of the proposed method MASFormer, the supervision signals, and the loss functions.

4.1. Two grouping modules in the MAS stage

The grouping operations in the MAS stage of the proposed method MASFormer aim to generate two groups according to each feature's view id and instance id.

The view ids and instance ids are predicted by the first FES stage in the proposed MASFormer, and the ground-truth ids are only used for supervising the predictions. The FES stage predicts the features F_{FES} for all instances from all views.

4.2. Two attention modules in the MAS stage

The mechanism of the two attention modules, including the view-level and instance-level attention modules, are the same. The difference exists in the meaning of the input and output. For example, for the view-level attention module in the MAS stage, the grouping operation takes all instances' features of the same view into the same group. Then, an attention map is created for each group, which describes the relationship among all instances' features. The values between relevant instances' features are higher than those between irrelevant instances' features in the attention map. The building of the attention map for the instance-level attention module is similar to the view-level attention module.

4.3. Supervision signals

For our proposed method and the baseline method, the supervision signals contain the amodal masks, category labels of all instances, instance ids, and view ids. The 3D models and the viewpoint information, including camera directions and positions, are *not* used.

4.4. Loss functions

The whole loss function consists of both the losses for amodal segmentation and the classification of each instance. For segmentation, the cross-entropy and dice losses are used to compute losses between predicted amodal masks and ground-truth amodal masks for all instances. For classification, the cross-entropy loss is computed between the predicted and ground-truth categories for all instances.

References

[1] Yi Hou, Chengyang Li, Yuheng Lu, Liping Zhu, Yuan Li, Huizhu Jia, and Xiaodong Xie. Enhancing and dissecting crowd counting by synthetic data. *ICASSP*, pages 2539–2543, 2022. 2

[2] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. *CVPR*, pages 4627–4635, 2017. 2