2D Amodal Instance Segmentation Guided by 3D Shape Prior

Zhixuan Li[1,2], Weining Ye[2], Tingting Jiang[✉1,2], and Tiejun Huang[2]

[1] Advanced Institute of Information Technology, Peking University, Hangzhou, China
[2] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing, China
{zhixuanli,ywning,ttjiang,tjhuang}@pku.edu.cn

Abstract. Amodal instance segmentation aims to predict the complete mask of the occluded instance, including both visible and invisible regions. Existing 2D AIS methods learn and predict the complete silhouettes of target instances in 2D space. However, masks in 2D space are only some observations and samples from the 3D model in different viewpoints and thus can not represent the real complete physical shape of the instances. With the 2D masks learned, 2D amodal methods are hard to generalize to new viewpoints not included in the training dataset. To tackle these problems, we are motivated by observations that (1) a 2D amodal mask is the projection of a 3D complete model, and (2) the 3D complete model can be recovered and reconstructed from the occluded 2D object instances. This paper builds a bridge to link the 2D occluded instances with the 3D complete models by 3D reconstruction and utilizes 3D shape prior for 2D AIS. To deal with the diversity of 3D shapes, our method is pretrained on large 3D reconstruction datasets for high-quality results. And we adopt the unsupervised 3D reconstruction method to avoid relying on 3D annotations. In this approach, our method can reconstruct 3D models from occluded 2D object instances and generalize to new unseen 2D viewpoints of the 3D object. Experiments demonstrate that our method outperforms all existing 2D AIS methods.

Keywords: Amodal, occlusion, instance segmentation

1 Introduction

Different from *visible* instance segmentation (VIS) [9, 1, 14, 31] which only predicts the visible region of each instance, *amodal* instance segmentation (AIS) [19, 37, 30] task poses a harder challenge that demands to predict both the visible and occluded parts. AIS has many potential applications, including auto-driving [27], automatic checkout in the market [7] and image editing [36].

The concept of AIS was proposed in 2016 [19], and several datasets [38, 5, 27, 13, 7] have been provided. Most of the existing *amodal* methods [19, 38, 5, 7, 13, 27, 34, 17] are developed based on *visible* instance segmentation methods [18, 9] that directly minimize the discrepancy between amodal prediction and ground-truth masks. Recently, some methods consider the characteristics of the amodal

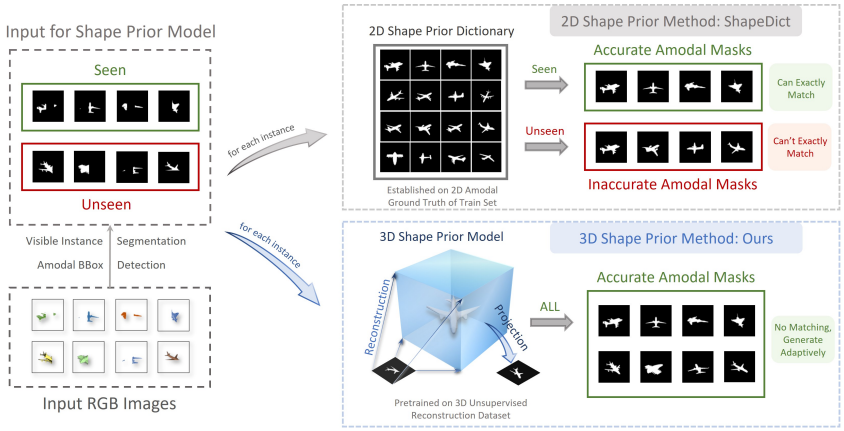

Fig. 1. Overview and comparison of the 2D shape prior dictionary (SPD) based method ShapeDict [30] and our proposed 3D shape prior generation-based method. For the left input RGB images, the first stage of both methods conduct instance segmentation to obtain *amodal* bounding boxes and *visible* masks for each instance. (a) ShapeDict regards each visible mask as *query* for SPD and *retrieves* the matched amodal shape prior masks. Due to the limited diversity of prestored shape prior, the retrieved shape prior is more appropriate for samples that *have been seen* (green box) in the dictionary rather than the *unseen ones* (red box). (b) Our proposed method adaptively generates 3D shape prior models from visible masks and performs projection for 2D amodal masks without needing for prestoring.

problem itself and propose new solutions. For example, relative depth order of instances is used to help comprehend the scene [37, 36]. Weakly supervised methods are proposed [23, 36, 25] without needing ground-truth amodal mask while taking ground-truth visible mask as input and supervision.

Besides, a natural solution is to use the shape-prior knowledge for handling the occluded instance, which lacks the shape and appearance information of the invisible region. In 2020, ShapeDict [30] proposes to utilize 2D shape prior knowledge to deal with the amodal segmentation problem. As shown in Fig. 1(a), ShapeDict first establishes a 2D shape prior dictionary (SPD) by applying the K-means algorithm on the ground-truth 2D amodal masks in the training set, and takes the cluster centers as shape priors. During inference, the closet shape prior to the predicted visible mask is retrieved from the SPD and used for amodal segmentation. However, this nearest neighbor search approach can only work for amodal masks having been seen during training (as shown in the *green* box of Fig. 1(a)). Otherwise, it will fetch inappropriate shape prior and lead to wrong amodal segmentation. For example, an occluded airplane photographed from a new viewpoint, whose amodal mask is not stored in the SPD, cannot be correctly matched. Therefore, ShapeDict is limited to the number and variety of shape prior masks stored in the dictionary, making it hard to generalize on unseen viewpoints. In this paper, we consider *if it is possible to adaptively generate the*

shape prior masks rather than prestoring the shape prior masks in a dictionary, and tackle the challenges of viewpoint changes?

With these problems in mind, we hope to learn the shape prior in the 3D space, which is a unified representation of 2D masks from all viewpoints and can generalize to new perspectives. Meanwhile, to avoid the shortcoming of the SPD method in ShapeDict, we hope to generate the 3D shape prior with learned shape knowledge adaptively and need no requirement for a prestored shape dictionary. To achieve these two purposes, we propose to reconstruct the complete 3D shape prior from the 2D occluded instance, as shown in Fig. 1(b). To accomplish 3D reconstruction, either multi-view images as input or 3D models as supervision signals are usually needed. Unfortunately, both are *not available* in any existing 2D AIS datasets. However, the good news is that, in recent years, single-view unsupervised 3D reconstruction methods [24, 20, 12] can avoid the requirements of multiple views and 3D model for training, which makes 2D amodal datasets usable for 3D reconstruction. For single-view unsupervised 3D reconstruction methods, the 3D model is first reconstructed from the single-view input RGB image and then projected along the estimated viewpoints to 2D masks. During reconstruction, only the 2D visible mask is provided as the supervision signal and the 3D reconstruction model is supervised *indirectly by the consistency between 2D projection of the 3D model and 2D visible mask*. In our method, the 2D amodal mask is used as the supervision signal for the single-view unsupervised 3D reconstruction.

In this work, we propose Amodal 3D Network (A3D), a novel coarse-to-fine architecture that combines category-specific 3D shape prior with 2D AIS. As shown in Fig. 2, for an input RGB image, we first apply the visible instance segmentation method to obtain visible masks of each instance. Next, we use a two-branch structure, in which the upper branch utilizes an *Encoder Decoder Network* for Category-specific 3D shape prior reconstruction, and the lower branch predicts camera viewpoint parameters by the *Viewpoint Estimator*. Then the *Differentiable Render* projects the 3D shape prior model according to the predicted viewpoint to the 2D coarse amodal mask. Finally, the *Region-specific Edge Refine module* refines the edges with the guidance of the visible mask and predicts the final amodal mask. With this coarse-to-fine pipeline, A3D can benefit from the power of 3D shape prior modelling and 2D edge refinement at the same time.

It is worth noting that our 3D shape prior reconstruction method *only requires 2D amodal masks as ground truth, without supervision signals like 3D models*, which are expensive to obtain. Because the 3D reconstruction module plays a crucial role in our method, we need to ensure that the reconstruction module can generate high-quality 3D shape prior models when facing 2D occluded instances. We design a pretrain-and-finetune pipeline that the single-view 3D reconstruction module is first pretrained on a large 3D reconstruction dataset like ShapeNet [2] for common shape *in unsupervised approach without using 3D annotations*. Then we conduct finetuning on the 2D AIS dataset for specific shape knowledge learning.

The effectiveness of our proposed method A3D is evaluated on several challenging datasets, including D2SA for market goods, KINS for person and vehicle, and COCOA-cls for life scene. We achieve state-of-the-art on all datasets.

We summarize our final contributions as follows:

1. A new method A3D is proposed for AIS, which utilizes the single-view unsupervised 3D reconstruction for 3D shape prior learning, to tackle the problem that 2D amodal segmentation methods are hard to generalize on new viewpoints. To our best knowledge, it is the first time the 3D shape prior is used for 2D AIS.
2. A coarse-to-fine pipeline is designed, which learns the 3D coarse shape prior and then refine edges with region-specific loss. It is end-to-end trainable and profits from both 3D and 2D information.

2 Related Work

2.1 2D Instance Segmentation

Amodal Instance Segmentation Comparing to visible instance segmentation, due to the shape of both visible and occluded regions needing to be predicted, the Amodal Instance Segmentation (AIS) task has fewer clues to infer the complete silhouette of instance and more ambiguity because of the occlusion. Existing methods mainly solve the task in 4 ways, including (1) directly minimizing between the prediction [19, 38, 5, 27] and the target, (2) using relative depth order to comprehend the relationship between different objects [37, 36], (3) mutual helping from visible and amodal masks [7, 17] and (4) using prestored shape prior knowledge [30, 23]. In the meanwhile, several datasets have been proposed including realistic ones [38, 27, 7] and synthetic ones [5, 13]. In addition, some papers [36, 23] are working on a relevant task *amodal completion*, which aims to predict the complete amodal shape based on the given visible mask, while there are no visible masks given in the AIS task.

All of the existing AIS algorithms are working in 2D space, which lacks comprehension of the real shape in 3D space. In this paper, our method learns the shape knowledge in 3D space to overcome the drawbacks of 2D AIS methods.

2.2 Unsupervised Single View 3D Reconstruction

Based on deep learning, supervised 3D reconstruction methods [4, 32, 33] are relying on high-quality 3D models for supervision, which are expensive to build. And because 2D supervision signals like segmentation masks are easier to obtain, unsupervised 3D model reconstruction methods are more popular. The pipeline of unsupervised 3D reconstruction consists of two steps, including 3D model reconstruction from the 2D image and rendering the 3D model into the 2D space, which is called *rasterization*. Whether the *rasterization* is differentiable decides whether the second rendering step can be included in the deep learning model with end-to-end training.

Before 2018, most methods [35, 15] take *rasterization by discrete assignment* and cannot be trained end-to-end for the whole network. To *make the rasterization differentiable*, in 2018, NMR [16] proposes an approximate gradient approach to make the backward gradient progress in rasterization differentiable. SoftRas [24] and DIB-R [3] methods make both of the forward and backward steps in rasterization differentiable and can be trained in an end-to-end manner. Based on the differentiable rasterization technique, UMR [20] utilizes the semantic parts consistency between 2D and 3D spaces as supervision. SMR [12] proposes landmark and interpolation consistency for self-supervision.

Because our method aims to represent and learn the complete shape in 3D space, we utilize the unsupervised single-view algorithm to gain a deeper understanding of the real shape in 3D space and help with 2D AIS with the reconstructed complete 3D shape.

3 Amodal 3D Network (A3D)

In this section, we develop a novel coarse-to-fine structure by combining the strength of 3D shape prior reconstruction and 2D edge refinement. We will first show the overall architecture of A3D and introduce each stage of A3D in detail.

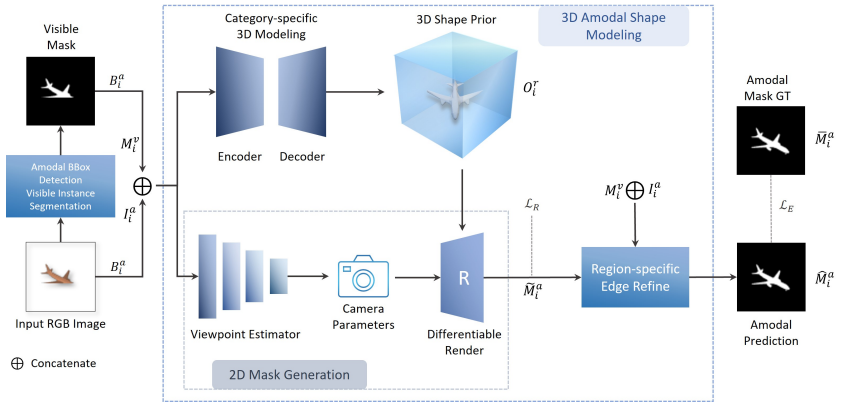

Fig. 2. The pipeline of our proposed Amodal 3D Network (A3D).

3.1 Overall Architecture

Given an input image $I \in R^{H \times W \times 3}$ containing N instances, for the i-th instance, 2D AIS algorithms aim to predict the class ID $c_i \in \{1, 2, ..., K\}$ and 2D amodal masks M_i^a.

The overall architecture is illustrated in Fig. 2. We take Amodal BBox Detection and Visible Instance Segmentation as the first stage to predict amodal

bounding box B_i^a and the visible mask M_i^v. Then the RGB image I_i^a is cropped by using B_i^a. For each instance, we concatenate M_i^v and I_i^a as input to the second proposed *3D Amodal Shape Modeling* (3D-ASM) stage for predicting the final amodal segmentation mask. 3D-ASM contains three important modules including *Category-specific 3D Modeling*, *2D Mask Generation* and *Region-specific Edge Refine*. (1) The *Category-specific 3D Modeling* module reconstructs the 3D complete model as shape prior. (2) In the *2D Mask Generation* module, the *Viewpoint Estimator* first predicts the camera parameters and apply the transformation to the reconstructed 3D shape prior model to the appropriate observation viewpoint. Then a *Differentiable Render* projects the 3D shape prior along the predicted viewpoint to obtain a coarse amodal segmentation mask. (3) Finally, the *Region-specific Edge Refine* module utilizes the visible mask, whose edge is accurate because the appearance information of the visible region is available to modify edges of the amodal mask. The whole network is end-to-end trainable.

We will introduce the details of our network in the following sections.

3.2 Amodal BBox Detection and Visible Instance Segmentation

In this stage, we aim to predict the amodal bounding boxes and the visible masks for each instance. Following [7, 37, 30] we choose the popular Mask-RCNN [9] method for this stage. In Mask-RCNN, the first *detection* stage is set to predict the *amodal bounding box* (BBox) and the second *segmentation* stage is set to predict *the visible mask*. For Mask-RCNN, we choose Faster-RCNN [28] for bounding box detection and ResNet-50 [11] combining Feature Pyramid Network [21] as the backbone network. We use the ground truth of *amodal bounding box*, *category ID* and *visible mask* as supervision signals. From this stage, we can obtain the predicted amodal bounding box B_i^a, class ID c_i, and visible foreground binary mask M_i^v in the region of amodal bounding box for the i-th instance. Then we crop the input RGB image I with amodal bounding box B_i^a to get the image I_i^a in the region of the i-th instance.

3.3 Category-specific 3D Modeling

In this module, we aim to reconstruct the complete 3D model based on the occluded instance in 2D space. To achieve the 3D reconstruction of the i-th occluded instance, traditional 3D reconstruction methods [4, 32, 33] requires either *multiple-view inputs* or *3D supervision signals*, which are not available in any existing 2D amodal segmentation datasets [7, 27, 38]. Therefore we choose to use the *single-view unsupervised* 3D reconstruction methods [24, 3, 20, 12] considering the dataset limitation.

In the single-view unsupervised 3D reconstruction framework, the 3D model is first reconstructed from 2D inputs (2D $\rightarrow$ 3D) and then projected to 2D space (3D $\rightarrow$ 2D) for 2D amodal mask predictions. The 3D reconstruction network is *indirectly* supervised by the ground-truth of the 2D amodal masks. There are not any 3D models used as supervision signals.

For the i-th instance, the input for 3D reconstruction is $A_i = [I_i^a, M_i^v]$, which is the concatenation of the image region I_i^a and visible mask M_i^v. We use the simple and classic *Encoder Decoder* structure following [16, 24] for 3D shape modeling, leaving room for improvement by using more complex models. The *Encoder* contains five *conv-bn-relu* blocks for visible feature extraction, and three *fully connected (fc)* layers for linearly feature mapping. There is also a *classification branch* taking the feature from Encoder output and predicts the category of each instance, making the 3D shape reconstruction in a class-specific manner. It is worthy to notice that we use one model to handle all categories.

We take a sphere as the initial object model $O_i^0(V_i^0)$, in which V_i^0 is the initial vertices. The *Decoder* consists of two *fc-relu* blocks to predict the offset $\triangle V_i$ between reconstructed 3D model and initial 3D model. Finally we can obtain the vertices $V_i^r = V_i^0 + \triangle V_i$ and the reconstructed 3D object $O_i^r(V_i^r)$. The detailed network architecture is described in the supplementary.

3.4 2D Mask Generation

For the i-th reconstructed 3D shape prior model O_i^r, if we want to obtain the 2D amodal mask, it is necessary to transform the 3D model with the correct parameters of the camera and project the 3D model to the camera plane. In this module, we take a *Viewpoint Estimator* to predict the camera parameters and utilize a *Differentiable Render* for projection.

Following SMR [12], we use an Encoder Network to construct the *Viewpoint Estimator*, which consists of five *conv-bn-relu* blocks and three *fully-connected* layers. The *Viewpoint Estimator* predicts the camera parameters $[e_i, d_i, (a_i^x, a_i^y)]$ representing elevation e_i, distance d_i and azimuth (a_i^x, a_i^y) in Cartesian coordinates, in which azimuth $a_i = arctan2(a_i^x, a_i^y)$. With the predicted viewpoint $[e_i, d_i, (a_i^x, a_i^y)]$, the 3D model is transformed appropriately. The *Viewpoint Estimator* is supervised *indirectly* by the ground-truth of 2D amodal masks, and no ground truth of viewpoints is used for supervision.

Finally to project the transformed 3D shape prior model O_i^r for the coarse 2D amodal mask $\widetilde{M}_i^a$, we utilize the *Differentiable Render* SoftRas [24], which can maintain the gradient flow for end-to-end training.

3.5 Region-specific Edge Refine

In previous modules, we have obtained the 3D reconstructed shape prior model O_i^r and the projected 2D coarse amodal mask $\widetilde{M}_i^a$. However, the quality of the existing 3D reconstruction method is affected by the number of vertices in the initial sphere O_i^0 and topological changes like holes which are hard to be learned in the mesh format. Therefore we design the *Region-specific Edge Refine* module to use the 2D amodal image region I_i^a and visible mask M_i^v to improve the edge of the amodal mask $\widetilde{M}_i^a$, because the appearance textures are only available in the visible region.

We take 3 repeated *conv-bn-relu* layers as the module architecture with kernel size=3. This module takes the concatenation of coarse 2D amodal mask $\widetilde{M}_i^a$, the

2D amodal image region I_i^a and visible mask M_i^v as input, and uses visible mask M_i^v to help with loss function, which is designed to punish more on the visible edge and less on the occluded edge. Visualization example of Edge Refine are shown in Fig. 3.

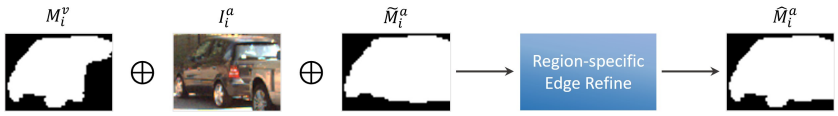

Fig. 3. Visualization of Region-specific Edge Refine. $\bigoplus$ means concatenation.

3.6 Loss Functions

In this section, we will introduce loss functions for our 3D modelling and the edge refinement modules.

3D Modeling To get rid of dependence on the expensive 3D model annotation, we choose to use the unsupervised 3D reconstruction method without needing 3D models as supervision signals. We train both the Category-specific 3D Modeling module and Region-specific Edge Refine module simultaneously because only the ground-truth 2D amodal masks $\overline{M}_i^a$ are available for supervision. Therefore the loss function is designed to encourage the predicted coarse amodal mask $\widetilde{M}_i^a$ being close to $\overline{M}_i^a$, which indirectly supervises the quality of reconstructed 3D shape prior model O_i^r. The loss function for unsupervised 3D reconstruction is:

$$\mathcal{L}_{\mathcal{R}} = \frac{1}{N}\sum_{i=1}^{N}(1 - IoU(\widetilde{M}_i^a, \overline{M}_i^a)) \tag{1}$$

where IoU computes the *intersection over union* between the predicted coarse amodal mask $\widetilde{M}_i^a$ and the ground-truth amodal mask $\overline{M}_i^a$.

Edge Refine In the *Region-specific Edge Refine* module, the loss function of the *Region-specific Amodal Edge Refine* module is designed as following:

$$\mathcal{L}_E = \frac{1}{N}\sum_{i=1}^{N}(\sum_{p\in S_i}\mathcal{L}_B(\widehat{M}_{i,p}^a, \overline{M}_{i,p}^a) + \lambda\sum_{p\notin S_i}\mathcal{L}_B(\widehat{M}_{i,p}^a, \overline{M}_{i,p}^a)) \tag{2}$$

where N is the instance number in the input image. For the i-th instance, S_i represents the visible region indicated by the ground-truth visible mask, and p denotes the pixel p. $\mathcal{L}_B$ is the *Binary Cross Entropy* loss function, computing the difference between the predicted values of pixel p from refined prediction $\widehat{M}_{i,p}^a$ and ground-truth mask $\overline{M}_{i,p}^a$. We set the loss weight $\lambda = 0.5$ to rely more on the visible region for edge refinement.

3.7 Pretrain and Finetune

In previous subsections, the whole framework of A3D and loss functions are introduced. In this subsection, a carefully designed pretrain and finetune strategy is introduced to improve the performance of 3D shape reconstruction for better 2D amodal segmentation results. It is worth noting that the ground truth of 3D models and viewpoints of the pretrain and finetune datasets are never used, to make our method applicable in real applications.

We first pretrain our A3D network on the 3D reconstruction dataset by unsupervised approach (as described in Sec 3.3), and then finetune on the train set of 2D AIS dataset, finally conduct inference on its test set.

To handle the problem that the categories of pretrain and finetune datasets are different, including overlapped and non-overlapped categories, we deal with them in different approaches. For overlapped categories, we reuse the network weights after pretraining as the initialization for training on the finetune dataset. With the weights being reused, the category-specific knowledge can be transfered for the overlapped categories between pretraining and finetune datasets. For non-overlapped categories, the weight parameters are randomly initialized [10]. The performance of overlapped and non-overlapped categories are shown in Tab. 3.

Besides, to improve the performance of 3D reconstruction supervised by single view, in the pretrain process, we use the cross-view technique [24], which requires the 3D model reconstructed from two viewpoints for the same object to be similar. The cross-view technique only additionally uses the correspondence information that two images from different viewpoints corresponds to the same object, and *never* takes the ground truth of viewpoints and 3D models as supervisions. The details are shown in the supplementary. The cross-view technique is *optional*.

4 Experiments

In order to evaluate our proposed method, extensive experiments have been conducted on three public amodal segmentation datasets, including COCOA-cls, KINS and D2SA, as well as a 3D dataset ShapeNet. Our method is compared with several SOTA amodal segmentation methods, and results show the advantage of our approach.

4.1 Datasets

2D AIS Datasets For 2D AIS task, we conduct experiments on a large-scale synthetic dataset ShapeNet [2] and three real 2D AIS datasets including COCOA-cls [7], KINS [27] and D2SA [7] for scenes of *outdoor, street and indoor supermarkets*. ShapeNet dataset contains 735,432 instances for training and 210,288 instances for testing. There are 13 categories, including various objects, and each object is rendered in 24 different viewpoints. We randomly occlude the RGB image and mask to simulate the occluded inputs. We do not

use any 3D shape and viewpoint supervision signals in all experiments even they are available. COCOA-cls dataset, which annotates a subpart of COCO [22] dataset with amodal masks, has 3,501 images and 10,592 instances in 80 categories. KINS dataset is the biggest street scene amodal dataset, which can be applied on tasks like auto driving, built on KITTI dataset [8] with re-annotated amodal masks. KINS has two super-classes, including person and vehicle, and seven sub-classes with 7,474 and 7,517 images for training and testing. D2SA dataset is built upon D2S [6] dataset with amodal mask re-annotated, including plenty kinds of goods placed in different postures and occlusion approaches on a rotatable supermarket platform with varying light conditions. D2SA contains 5,600 images totally and 28,720 instances in 60 classes.

Pretraining Datasets For using the pretrain and finetune strategy claimed in Sec. 3.7, both ShapeNet and PASCAL3D+ [29] are used. PASCAL3D+ dataset contains 55,867 3D models in 12 categories, which is more than 39,405 3D models in ShapeNet dataset, providing richer shape knowledge for 3D reconstruction.

4.2 Implementation Details and Evaluation Metric

Our method is implemented based on the Pytorch [26] framework. For all of the 2D amodal segmentation methods used in our experiments, we use the same configuration following ShapeDict [30]. For our proposed A3D Network, the learning rate is set to 0.0001, and we use the Adam algorithm for gradient descent with 64 batches. All experiments are conducted on a single 2080Ti GPU card. All categories including *rigid and non-rigid* are used in all experiments. Faster-RCNN with ResNet-50 is used for all methods for object detection.

We choose the *mean Intersection over Union* (mIoU) and *mean Average Precision* (mAP) as metrics for performance evaluation. It is worthy to notice that the commonly chosen metric *mean Average Precision* (mAP) measures the performance of two sub-tasks in *Amodal Instance Segmentation* simultaneously, including *Object Detection* and *Semantic Segmentation*. Therefore for ShapeDict dataset we only report mIoU because there are only one object in each image and mAP which measures the performance of object detection is not reported.

4.3 2D Amodal Instance Segmentation

This section evaluates 2D AIS methods on the challenging ShapeNet dataset and three amodal datasets, including COCOA-cls, D2SA and KINS for different scenes. Following state-of-the-art methods are used for comparison.

(1) Mask-RCNN [9] is trained using ground-truth *amodal* bounding boxes and masks to show the transferability of the visible instance segmentation method on the amodal problem. (2) ORCNN [7] is a two-branch approach that predicts and supervises the visible, amodal and occluded region at the same time. (3) BCNet [17] decouples the occluding and occluded instances combining graph

convolution. (4) ShapeDict [30] establishes a 2D shape prior dictionary by clustering the ground-truth amodal masks and uses the *query-and-retrieve* approach to provide prior knowledge. Besides, we also take Deocclusion [36], a weakly-supervised amodal completion method, for comparision to show the performance gap between weakly and fully supervised AIS methods.

Table 1. Results (mIoU) on the ShapeNet dataset. For each category, bold performance is the best, and the second-best is underlined. The subscript numbers are the subtraction results between ours and the second-best methods. SU means the Supervision signal type. W and F mean weakly and fully supervised.

Methods	SU	Airplane	Bench	Dresser	Car	Chair	Display	Lamp	Speaker	Rifle	Sofa	Table	Phone	Vessel	mIoU
Deocclusion [36]$_{\text{CVPR'20}}$	W	24.9	67.4	45.3	58.8	83.7	78.4	77.9	15.2	48.7	48.1	39.5	23.8	71.9	52.2
Mask-RCNN [9]$_{\text{ICCV'17}}$	F	73.4	66.0	92.4	93.5	**89.3**	**90.0**	77.4	88.5	30.0	86.0	73.1	89.8	80.5	79.2
ORCNN [7]$_{\text{WACV'19}}$	F	71.5	61.1	92.0	92.7	88.8	88.8	79.5	88.7	32.8	85.6	72.5	89.0	80.0	78.7
BCNet [17]$_{\text{CVPR'21}}$	F	73.0	75.1	93.8	89.4	86.6	88.7	**81.6**	90.2	32.8	83.4	77.5	88.7	74.8	78.2
ShapeDict [30]$_{\text{AAAI'21}}$	F	75.2	68.5	93.7	**93.6**	88.4	89.3	78.1	88.6	34.4	87.3	74.8	90.7	80.9	80.3
Ours (no pretrain)	F	**77.9**	**80.8**	**94.2**	92.8	79.7	87.5	67.8	**90.5**	**69.9**	**90.3**	**86.2**	**92.1**	**81.3**	**83.9**

Table 2. Results (mIoU and mAP) on the 2D AIS datasets. SU means the Supervision signal type. W and F mean weakly and fully supervised. FLOPs and Params measure the computational efficiency and model size.

Method	SU	mIoU ↑			mAP ↑			FLOPs(G) ↓	Params(M) ↓
		D2SA	KINS	COCOA-cls	D2SA	KINS	COCOA-cls		
Deocclusion [36]$_{\text{CVPR'20}}$	W	73.8	59.2	39.2	61.7	27.5	19.9	160.4	44.1
Mask-RCNN [9]$_{\text{ICCV'17}}$	F	74.6	60.1	63.8	63.6	30.0	33.7	160.4	44.1
ORCNN [7]$_{\text{WACV'19}}$	F	74.1	55.1	57.6	64.2	30.6	28.0	229.4	46.8
BCNet [17]$_{\text{CVPR'21}}$	F	74.9	44.0	15.1	50.9	22.1	16.2	263.5	63.2
ShapeDict [30]$_{\text{AAAI'21}}$	F	75.0	63.7	64.5	70.3	32.1	35.4	271.3	48.0
Ours (w/o pretrain)	F	74.7	61.4	64.2	68.5	31.4	34.9	229.8	57.2
Ours (w/ pretrain)	F	**78.4**	**65.5**	**67.4**	**73.5**	**36.2**	**40.6**	229.8	57.2

The comparison results of AIS methods are shown in Tab. 1 for ShapeNet dataset and Tab. 2 for the three 2D AIS datasets, including D2SA, KINS and COCOA-cls. In Tab. 1, all methods are trained on the train set of ShapeNet and there are no extra data used for pretraining in our method. Our method achieves the best performance on nine categories. Compared with the methods of the second-best performance, A3D gains significant IoU improvement on *Bench, Rifle and Table* with 5.7%, 21.6% and 8.7% respectively, which shows the effectiveness of our proposed A3D method. Fig. 4 shows the qualitative results of our A3D method on the ShapeNet dataset.

In Tab. 2, all methods except our methods are trained on the train split of respective datasets, and our method additionally uses ShapeNet and PASCAL3D+ for pretraining to make our method applicable in real applications. Our A3D network outperforms all methods with certain advantages for mIoU and mAP. In terms of the number of FLOPs and Parmeters, our method is com-

parable to previous work. More visualizations for D2SA, KINS and COCOA-cls datasets are given in the supplementary.

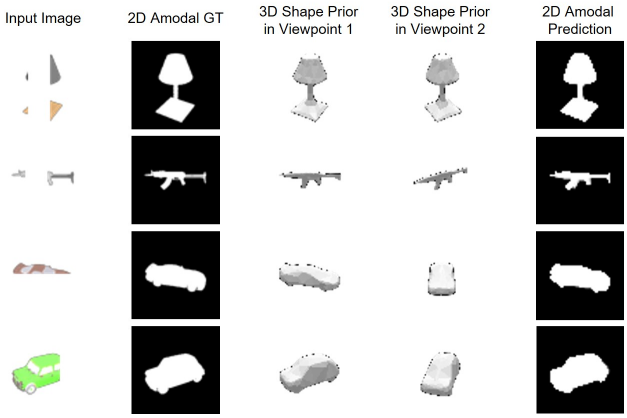

Fig. 4. Visualization result of our method on the ShapeNet dataset. The reconstructed 3D shape prior models are shown from two viewpoints.

4.4 Effectiveness of Pretraining

Table 3. Ablation study results (mAP) of pretraining. N, S, P and S+P means no pretraining, pretraining with ShapeNet, with PASCAL3D+, and with both ShapeNet & PASCAL3D+. #Overlapped means the number of categories overlapped between pretrain and finetune datasets. SU means supervision signal (W and F for weakly and fully supervised). Category number is noted in the brackets after each dataset name.

Index	Methods	SU	D2SA (60)				KINS (7)				COCOA-cls (80)			
			N	S	P	S+P	N	S	P	S+P	N	S	P	S+P
#Overlapped			-	2	2	4	-	1	4	4	-	13	12	19
1	Deocclusion	W	61.7	61.9	62.1	62.3	27.5	27.9	28.2	28.8	19.9	20.4	20.9	21.3
2	Mask-RCNN	F	63.6	63.9	64.2	64.8	30.0	30.5	30.8	31.1	33.7	33.9	34.2	34.6
3	ORCNN	F	64.2	64.8	65.1	65.5	30.6	30.9	31.2	31.8	28.0	28.3	28.7	29.1
4	BCNet	F	50.9	51.2	51.5	51.9	22.1	22.6	22.8	23.0	16.2	16.8	17.1	17.4
5	ShapeDict	F	**70.3**	70.5	70.9	71.2	**32.1**	32.2	32.4	32.5	**35.4**	35.7	36.1	36.4
6	Ours	F	68.5	**71.4**	**72.6**	**73.5**	31.4	**33.7**	**35.2**	**36.2**	34.9	**37.4**	**39.2**	**40.6**

In Tab. 2, all the previous methods do not use the pretrain dataset while our method does. To make a fair comparison, we design an experiment such that each previous method can also take advantage of the pretrain dataset. Specifically, in pretrain process, each previous method can take all images in the pretrain dataset, each of which contains one non-occluded object with white background

and the corresponding 2D amodal mask, as training data. With this pretrain strategy, all the previous methods can also make use of the pretrain dataset and thus the comparison between previous methods and ours is fair. We compare the performance of all methods with & without pretraining on three 2D AIS datasets. Results are shown in Tab. 3.

Comparing methods in different lines, we can conclude that by directly adding the 3D representation without pretraining, our method outperforms the baseline method Mask-RCNN but cannot beat ShapeDict. This is because the 2D AIS datasets provide not enough supervision for training 3D reconstruction in our method. In the last line of our method, for each finetune dataset, the performance increases with more pretrain data used. Meanwhile the performance of 2D AIS methods (Tab. 3, Line #1 to #5) do not increase much with pretraining. This is because for both pretrain datasets, each input RGB image contains only one non-occluded object with white background, which is easy to be segmented and not very helpful for amodal segmentation. However the pretrain datasets are very helpful for 3D reconstruction, making our method gains much improvements. With limited number of overlapped categories between pretrain and finetune datasets, our method can still achieve good performances.

4.5 Ablation Study

In this section, we conduct ablation experiments to validate the effectiveness of our proposed modules and pretraining for 3D reconstruction.

Effectiveness of 3D Modeling and Edge Refine In our proposed A3D network, we design a Category-specific 3D Modeling module for 3D shape prior generation and a Region-specific Edge Refine module for 2D edge refinement. In this section, we validate the effectiveness of the two proposed modules on the ShapeNet dataset. (1) The Mask-RCNN method, which directly predicts the 2D amodal mask from the input image, is the baseline method. (2) If the Category-specific 3D Modeling module is added, 4% mIoU improvement will be obtained, which shows the effectiveness of 3D modeling. (3) Then after further combining the Region-specific Edge Refine module, the performance can boost 0.7% mIoU, which improves the quality of some details of the edge and not drastically modifies the predicted mask. Visualizations of the effectiveness are shown in Fig. 3, and Edge Refine can improve the quality of boundary to some extent.

Effectiveness of cross-view technique Cross-view technique supervises the reconstructed 3D model from two viewpoints to be consistent. It does not use any additional supervision signals like viewpoint information or 3D model, only use the correspondence of two images from the same object of different viewpoints. We evaluate the performance without and with cross-view technique on ShapeNet dataset, and the mIoU results are 82.5% and 83.9% respectively. Cross-view technique brings 1.4% improvement with the reconstruction consistency between different viewpoints.

4.6 Methodology Limitation

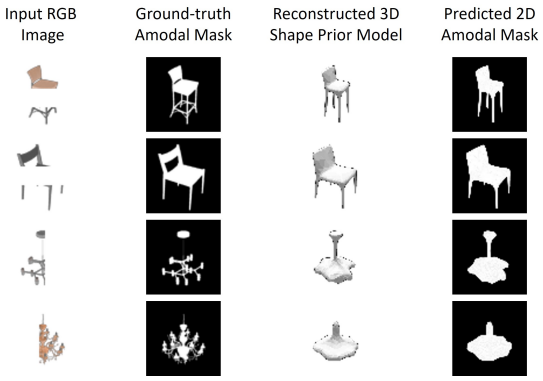

Fig. 5. Examples of chairs and lamps. In each four-tuple, images from left to right are input RGB images, ground-truth amodal masks, reconstructed 3D shape prior models, and predicted amodal masks.

As shown in Tab. 1, for the categories of *Chair* and *Lamp*, ours A3D method fails to perform well with large margin, dropping for 9.6% and 11.7% IoU compared with the best performance. As illustrated in Fig. 5, there are plenty of holes in the *Chair* category and complicated structures in the *Lamp* category. However, in our A3D method, the 3D shape prior model is reconstructed by predicting vertices offset from the initial sphere, remaining the topology unchanged. Complicated structures in both *Chair* and *Lamp* categories require the topology changes, where our A3D network is incapable at present. We leave this problem to future work.

5 Conclusion

In this paper, we propose a novel coarse-to-fine Amodal 3D (A3D) network. A3D is a brand new framework which for the first time tackles the 2D AIS problem by reconstructing the 3D complete shape prior model. With the benefits of 3D modelling, A3D can alleviate the shortcoming that 2D AIS methods are difficult to generalize on untrained new viewpoints of the occluded 3D object. Our A3D achieves state-of-the-art performance on multiple AIS datasets.

Acknowledgments

This work was partially supported by the Natural Science Foundation of China under contracts 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. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: Yolact: Real-time instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9157–9166 (2019)
2. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
3. Chen, W., Ling, H., Gao, J., Smith, E., Lehtinen, J., Jacobson, A., Fidler, S.: Learning to predict 3d objects with an interpolation-based differentiable renderer. Advances in Neural Information Processing Systems **32**, 9609–9619 (2019)
4. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: European conference on computer vision. pp. 628–644. Springer (2016)
5. Ehsani, K., Mottaghi, R., Farhadi, A.: Segan: Segmenting and generating the invisible. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6144–6153 (2018)
6. Follmann, P., Bottger, T., Hartinger, P., Konig, R., Ulrich, M.: Mvtec d2s: Densely segmented supermarket dataset. In: Proceedings of the European conference on computer vision (ECCV). pp. 569–585 (2018)
7. Follmann, P., König, R., Härtinger, P., Klostermann, M., Böttger, T.: Learning to see the invisible: End-to-end trainable amodal instance segmentation. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1328–1336. IEEE (2019)
8. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. 2015 IEEE International Conference on Computer Vision (ICCV) pp. 1026–1034 (2015)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Hu, T., Wang, L., Xu, X., Liu, S., Jia, J.: Self-supervised 3d mesh reconstruction from single images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6002–6011 (2021)
13. Hu, Y.T., Chen, H.S., Hui, K., Huang, J.B., Schwing, A.G.: 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. pp. 3105–3115 (2019)
14. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring r-cnn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6409–6418 (2019)
15. Kanazawa, A., Tulsiani, S., Efros, A.A., Malik, J.: Learning category-specific mesh reconstruction from image collections. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 371–386 (2018)
16. Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3907–3916 (2018)

17. Ke, L., Tai, Y.W., Tang, C.K.: Deep occlusion-aware instance segmentation with overlapping bilayers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4019–4028 (2021)
18. Li, K., Hariharan, B., Malik, J.: Iterative instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3659–3667 (2016)
19. Li, K., Malik, J.: Amodal instance segmentation. In: European Conference on Computer Vision. pp. 677–693. Springer (2016)
20. Li, X., Liu, S., Kim, K., De Mello, S., Jampani, V., Yang, M.H., Kautz, J.: Self-supervised single-view 3d reconstruction via semantic consistency. In: European Conference on Computer Vision. pp. 677–693. Springer (2020)
21. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
23. Ling, H., Acuna, D., Kreis, K., Kim, S.W., Fidler, S.: Variational amodal object completion. In: NeurIPS (2020)
24. Liu, S., Li, T., Chen, W., Li, H.: Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7708–7717 (2019)
25. Nguyen, K., Todorovic, S.: A weakly supervised amodal segmenter with boundary uncertainty estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7396–7405 (2021)
26. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019)
27. Qi, L., Jiang, L., Liu, S., Shen, X., Jia, J.: Amodal instance segmentation with kins dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3014–3023 (2019)
28. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**, 91–99 (2015)
29. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond pascal: A benchmark for 3d object detection in the wild. In: IEEE winter conference on applications of computer vision. pp. 75–82. IEEE (2014)
30. Xiao, Y., Xu, Y., Zhong, Z., Luo, W., Li, J., Gao, S.: Amodal segmentation based on visible region segmentation and shape prior. arXiv preprint arXiv:2012.05598 (2020)
31. Xie, E., Sun, P., Song, X., Wang, W., Liu, X., Liang, D., Shen, C., Luo, P.: Polarmask: Single shot instance segmentation with polar representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12193–12202 (2020)
32. Xie, H., Yao, H., Sun, X., Zhou, S., Zhang, S.: Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2690–2698 (2019)

33. Xie, H., Yao, H., Zhang, S., Zhou, S., Sun, W.: Pix2vox++: Multi-scale context-aware 3d object reconstruction from single and multiple images. International Journal of Computer Vision **128**(12), 2919–2935 (2020)
34. Yan, X., Wang, F., Liu, W., Yu, Y., He, S., Pan, J.: Visualizing the invisible: Occluded vehicle segmentation and recovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7618–7627 (2019)
35. Yan, X., Yang, J., Yumer, E., Guo, Y., Lee, H.: Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. arXiv preprint arXiv:1612.00814 (2016)
36. Zhan, X., Pan, X., Dai, B., Liu, Z., Lin, D., Loy, C.C.: Self-supervised scene de-occlusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3784–3792 (2020)
37. Zhang, Z., Chen, A., Xie, L., Yu, J., Gao, S.: Learning semantics-aware distance map with semantics layering network for amodal instance segmentation. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 2124–2132 (2019)
38. Zhu, Y., Tian, Y., Metaxas, D., Dollár, P.: Semantic amodal segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1464–1472 (2017)