Amodal Ground Truth and Completion in the Wild

Guanqi Zhan[1], Chuanxia Zheng[1], Weidi Xie[1,2], Andrew Zisserman[1]
[1]VGG, University of Oxford, [2]CMIC, Shanghai Jiao Tong University
{guanqi,cxzheng,weidi,az}@robots.ox.ac.uk

Abstract

This paper studies amodal image segmentation: predicting entire object segmentation masks including both visible and invisible (occluded) parts. In previous work, the amodal segmentation ground truth on real images is usually predicted by manual annotaton and thus is subjective. In contrast, we use 3D data to establish an automatic pipeline to determine authentic ground truth *amodal masks for partially occluded objects in real images. This pipeline is used to construct an amodal completion* evaluation *benchmark,* MP3D-Amodal*, consisting of a variety of object categories and labels. To better handle the amodal completion task in the wild, we explore two architecture variants: a two-stage model that first infers the occluder, followed by amodal mask completion; and a one-stage model that exploits the representation power of Stable Diffusion for amodal segmentation across many categories. Without bells and whistles, our method achieves a new state-of-the-art performance on Amodal segmentation datasets that cover a large variety of objects, including COCOA and our new* MP3D-Amodal *dataset. The dataset, model, and code are available at* https://www.robots.ox.ac.uk/~vgg/research/amodal/.

1. Introduction

The vision community has rapidly improved instance segmentation performance over the last few years with a succession of powerful models, such as Mask-RCNN [13], Mask2Former [5], and Seg-Anything (SAM) [23]. However, despite this remarkable progress, these models only provide pixel-level *modal* segmentations for objects in the images, *i.e.*, the instance masks are for the *visible* pixels. The models lack the human ability to infer the full extent of the object in an image, when it is partially occluded. The prediction of *amodal masks*, which covers the full extent of the object, can assist several downstream tasks including object detection [48], smart image editing [29, 47, 51], 3D reconstruction from a single image [17, 18, 34, 44, 55], object permanence in video segmentation [14, 41, 46], predicting support relationships between objects [38, 54], and

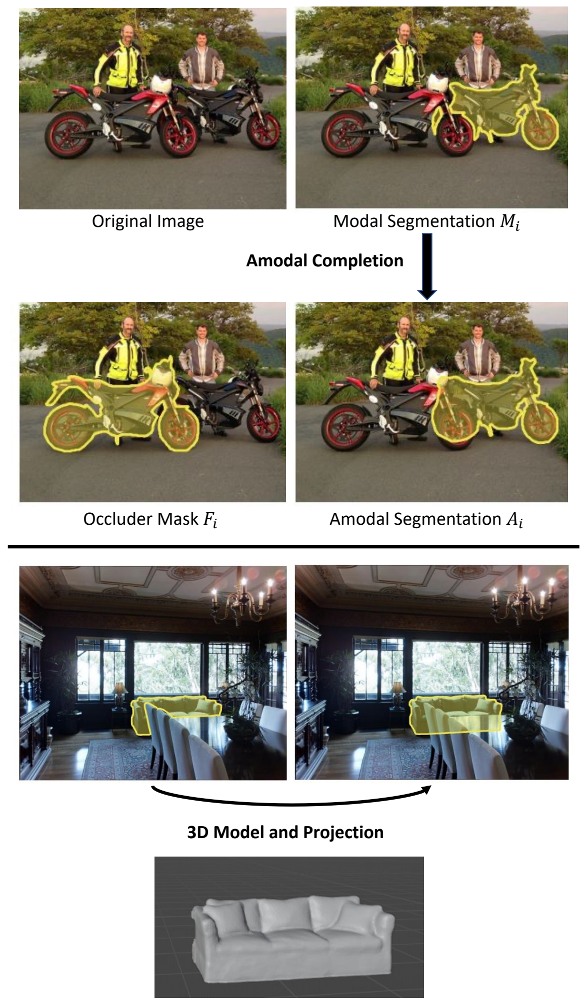

Figure 1. **Amodal Ground Truth and Completion in the Wild.** Top: The task of amodal completion is to predict the amodal mask A_i for an object 'i' in the image specified by the modal mask M_i (here the object is the rear motorbike). Previous methods [31, 51] require the mask of the occluder F_i to be also provided to do the task; but our goal is to predict the amodal mask when the occluder mask is *not* provided and the occluded object is of *any category*. Bottom: We propose a novel method for generating amodal masks for real images: using 3D structure to produce *ground truth* modal and amodal masks for object instances. The method is used to generate a ground truth evaluation dataset for real images.

Dataset	Amodal GT	Image Type	# Categories	# Images	Type
COCOA [53]	✗	Real	Multiple	5,073	General
COCOA-cls [9]	✗	Real	80	3,499	General
KINS [35]	✗	Real	8	14,991	Vehicle
DYCE [8]	✓	Synthetic	79	5,500	Indoor
OLMD [7]	✓	Synthetic	40	13,000	Indoor
CSD [52]	✓	Synthetic	40	11,434	Indoor
D2SA [9]	✓	Synthetic	60	5,600	Industrial
KITTI-360-APS [30]	✗	Real	17	61,168	Vehicle
BDD100K-APS [30]	✗	Real	16	202	Vehicle
WALT [37]	✓	Real	2	60,000	Vehicle
MUVA [27]	✓	Synthetic	20	26,406	Shopping
MP3D-Amodal (Ours)	✓	Real	427(40)	10,883	Indoor

Table 1. **Comparison of Different Amodal Datasets**. Amodal GT: whether the dataset provides ground truth amodal annotations or is manually guessed. # represents the number of the following types. Our MP3D-Amodal dataset (Sec. 3) has 427 different semantic labels mapped to 40 different MatterPort categories. Note, the WALT dataset consists of video frames from 12 camera viewpoints, mainly of vehicles moving.

more generally for planning in manipulation and navigational tasks where reasoning on the full extent of the object could lead to improvements [16, 20, 21, 42, 43].

Predicting amodal masks for objects in 2D images is challenging: this is because real scenes contain a vast collection of different types of objects, resulting in complex occlusions when they are projected to 2D images from a 3D physical world. To accurately complete the amodal shape of partially occluded objects requires a prediction of occlusion relations (in order to infer if and where the object is partially occluded), as well as predicting the shape of the occluded regions. This challenge is also reflected in the type of amodal datasets available which are mostly synthetic – for real images, amodal masks are almost always 'imagined' by human annotators providing the contour of the amodal mask, and there is no dataset available to evaluate amodal completions with *authentic* ground truth annotations for a large variety of object categories (see Table 1).

In this paper, our first contribution is to provide a new amodal benchmark evaluation dataset based on authentic ground truth amodal segmentation for real images, and covering a large variety of objects. The new dataset is called *MP3D-Amodal*, and the amodal mask is obtained by projecting the 3D structure of objects in the scene onto the image (Figure 1 bottom). We build the dataset from MatterPort3D [3] that has both 3D structure and real images of indoor scenes. The dataset and generation method is described in Sec. 3.

In most prior work, amodal completion algorithms required the occluder mask to be specified [31, 51] (Figure 1 top). Our second contribution is to propose two architecture variants that do not require the occluder mask to be supplied: *OccAmodal*, a two-stage model that first infers the occluder, followed by amodal mask compeletion; and *SDAmodal*, a one-stage model that uses the features of a pre-trained Stable Diffusion network, benefiting from its strong outpainting capabilities. These architectures are described in Sec. 4.

We achieve state-of-the-art amodal completion performance on both the public COCOA [53] dataset, and on our own *MP3D-Amodal* benchmark. In particular, the one-stage model, *SDAmodal*, benefiting from the pre-trained Stable Diffusion model, is able to generalize to another dataset with objects from a different domain and different categories, demonstrating class-agnostic completion. Taken together, the handling of situations where occluder masks are not provided and the class-agnostic domain generalization, moves amodal completions towards an 'in the wild' capability.

2. Related Work

Amodal Datasets. In the literature, there have been continuous efforts on creating datasets for amodal segmentation, for example, COCOA [9, 53] builds on COCO [28], KINS [35] builds on KITTI [11]. However, the ground truth amodal masks for both of these datasets are created based on the 2D images, thus inevitably requiring human imagination for the occluded regions. To improve the quality of ground truth amodal mask, the DYCE [8], OLMD [7], CSD [52] and MUVA [27] datasets were created by rendering the whole scene and corresponding individual intact objects using synthetic 3D models. WALT [37] collected objective amodal masks via time-lapse imagery, but their objects are limited to cars and humans. Table 1 provides a summary of the datasets currently available. In contrast, we are the first to collect a complex dataset that provides authentic amodal ground truth for the occluded objects of a large variety of categories in real scenes.

Amodal Instance Segmentation. Classical instance segmentation methods [1, 4, 5, 13, 23, 32] mainly focus on segmenting *visible* pixels, while amodal instance segmentation [24] aims to detect the objects as a whole, *i.e.*, both *visible* and *invisible* parts. These methods are usually trained on images [9, 10, 19, 25, 26, 30, 35, 39, 40, 45, 52] with

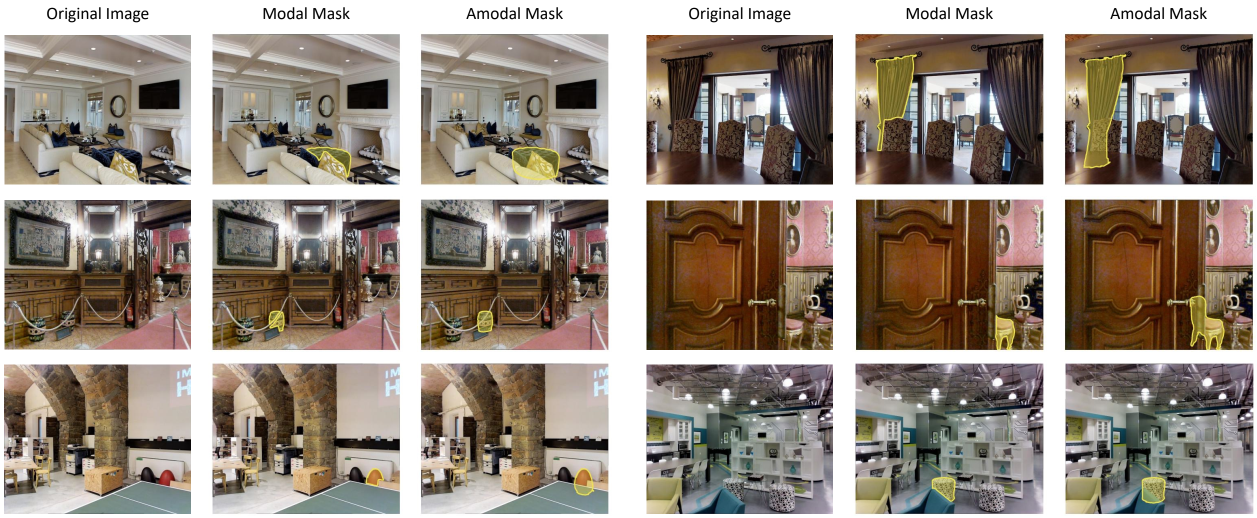

Figure 2. **Samples from the *MP3D-Amodal* Dataset.** For each sample, the original image is shown together with the generated modal and amodal masks.

manually annotated ground truth amodal masks in a fully-supervised manner. However, these methods are trained on datasets with limited number of object classes, *e.g.*, 80 categories for COCOA-cls, and are class dependent.

Amodal Completion is conceptually similar to amodal instance segmentation, except that here the *visible* mask for the target object is already provided as input. Most existing methods [31, 51] assume the occluder mask is provided and cannot handle the situation where the object is occluded by an unknown occluder, *i.e.*, the occluder mask is not provided or the occluder is difficult to define. The methods are trained on COCOA, which covers a large variety of categories, and are more class-agnostic than methods trained on COCOA-cls. Another work [29] uses VAE [22] to model the task of amodal completion, but can only handle limited categories in driving scenes.

3. The MP3D-Amodal Dataset

In this section we describe the new amodal dataset *MP3D-Amodal*, that is constructed from the MatterPort3D [3] dataset. We first overview the contents of the dataset in Sec. 3.1, and then describe our method of generating ground truth amodal masks on real images from 3D data in Sec. 3.2.

3.1. An Overview of the Dataset

The dataset contains 12,724 annotated amodal ground truth masks for over 10,883 real images. Since it is built from the MatterPort dataset, we use the classifications inherited from that dataset, where objects are described by their *category* and *semantic labels*. Note that, categories are more coarse-grained than semantic labels and one category may contain several different semantic labels, *e.g.*, the category 'chair' contains semantic labels 'dining chair', 'sofa chair' and 'armchair'; and the category 'appliances' contains 'refrigerator', 'oven', and 'washing machine'.

Split	# Scenes	# Images	# Instances	# MatterPort Categories	# Semantic Labels
Training	4	1,100	1,283	35	130
Evaluation	86	9,783	11,441	40	385
Total	90	10,883	12,724	40	427

Table 2. **Statistics of the generated *MP3D-Amodal* dataset.** Each instance has a semantic label as annotated in the MatterPort3D dataset, which is also mapped to a more general MatterPort category. Across the training and evaluation splits, there are 88 semantic labels in common, and 297 semantic labels in the evaluation split but not in the training split.

Table 2 gives the details for dataset splits. To have a better and more comprehensive evaluation, we make the evaluation split to have more scenes than the training split. Across the training and evaluation splits, there are 88 semantic labels in common, and 297 semantic labels in the evaluation split but not in the training split. A small part of the collected dataset is reserved for training, as this allows some domain adaptation for an algorithm. The scenes of the training set are disjoint from those of the evaluation set.

Samples from the dataset are displayed in Figure 2. The dataset contains diverse range of objects, with some categories not in the 'general' COCOA dataset, *e.g.,* the example in the bottom left of Figure 2 is a novel category. More examples of the dataset are in the ArXiv version of this paper [50].

Figure 4 visualizes the distributions of the dataset in terms of the number of instances for each MatterPort category, and the number of instances for different occlusion ratios,

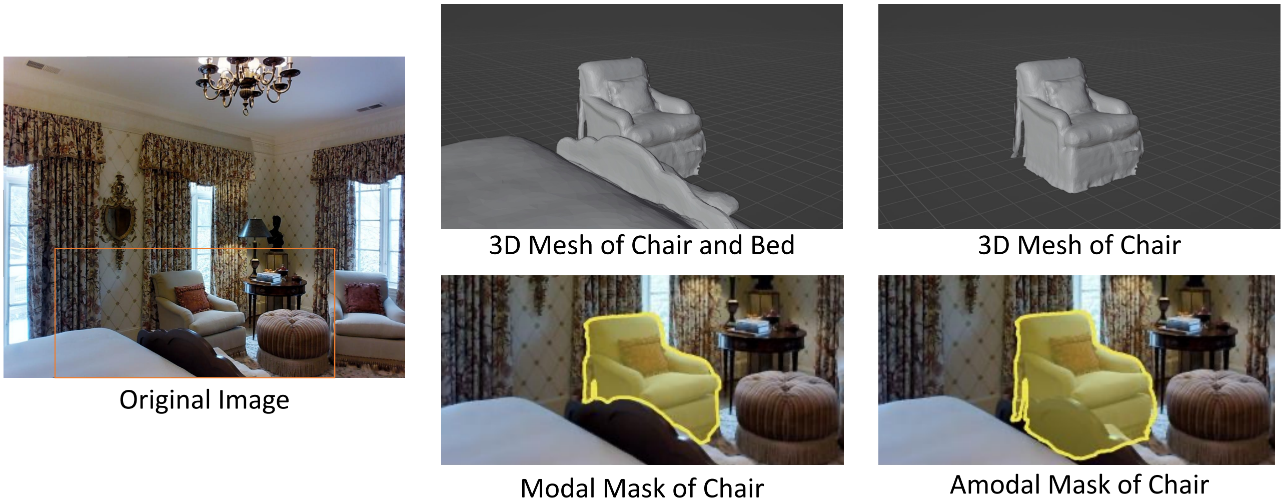

Figure 3. **Automated Generation of the *MP3D-Amodal* Ground Truth Dataset.** The dataset is automatically generated from the MatterPort3D [3] dataset, and provides ground truth modal and amodal masks for objects in real images. The generation process is illustrated here for the chair and proceeds in two steps: first, modal and amodal masks in a particular image are obtained for each object by projecting the object's 3D mesh individually (for the amodal mask), and also by projecting the 3D mesh of all objects (for the modal mask). In this example, the 3D mesh of the bed occludes the chair when projected into the image. In the second step, an object is selected for the dataset if its amodal mask is larger than its modal mask by a threshold. In this case the chair is selected, but other objects such as the stool would not be selected since it is not occluded by other objects in this viewpoint, and so their modal and amodal masks would be the same.

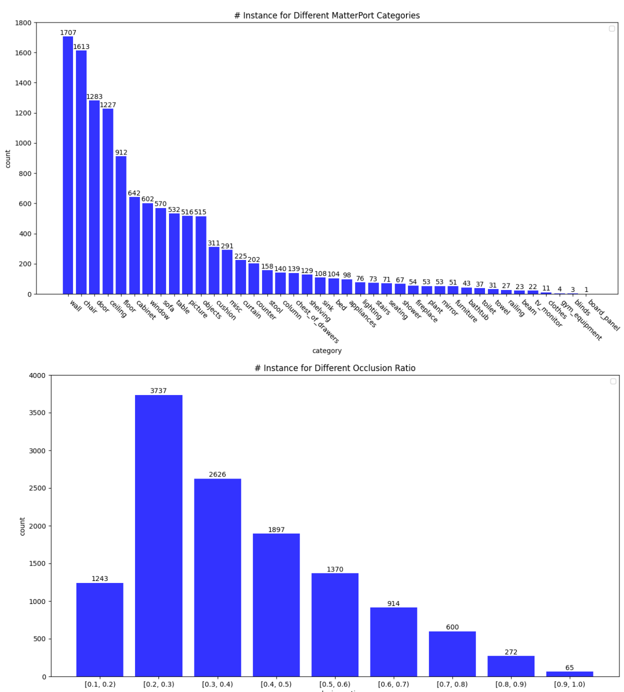

Figure 4. **Distributions of the *MP3D-Amodal* Dataset** in terms of the number of instances for each MatterPort category, and the number of instances for different occlusion ratios.

where the occlusion ratio is the proportion of the the object that is occluded (the difference between amodal and modal masks, divided by the area of the amodal mask). It is evident that there is a wide range of occlusion ratios, from slightly occluded to severely occluded.

3.2. Generating Amodal Ground Truth from 3D

We exploit the MatterPort3D [3] dataset, that is equipped with two essential elements: a 3D mesh for each object instance in the scene, and real images (and associated cameras) of the scene. In the following we detail the procedure for automatically generating amodal and modal masks of individual objects. The process is illustrated in Figure 3.

Modal Mask Generation. For a particular scene, we obtain 2D instance segmentations (a modal mask for each object) by projecting all objects with their instance labels together onto the image with the associated camera. If M_i and O_i denote the modal mask and 3D mesh of the i-th object, Φ refers to the projection of 3D meshes to the camera space, then the modal masks of the image are:

$$\{M_1, M_2, \ldots, M_n\} = \Phi(O_1 \cup O_2 \cup \cdots \cup O_n) \quad (1)$$

Amodal Mask Generation. The amodal mask A_i for each object i is simply obtained by projecting each object to the camera separately:

$$\{A_i\} = \Phi(O_i), \qquad \forall i \in \{1, 2, 3, \ldots, n\} \quad (2)$$

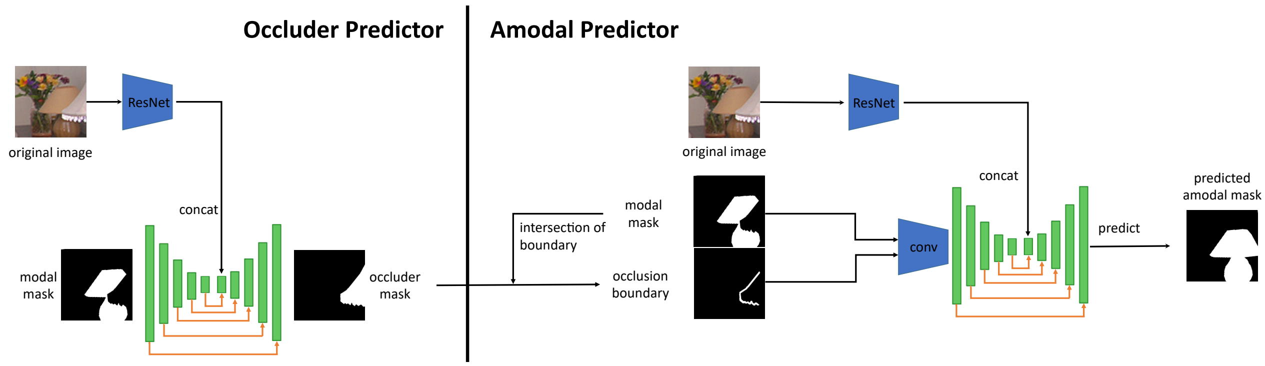

Figure 5. **Two-Stage Architecture (OccAmodal) for Amodal Prediction.** *Left*: A lightweight U-Net based architecture is used to predict the occluder mask for each object. *Right*: The amodal predictor takes the predicted occluder mask, together with the modal mask and image as input to predict the amodal segmentation mask.

Occluded Object Selection. Then partially occluded objects are identified and selected by comparing the modal and amodal mask. If the amodal mask of the object is larger than the modal mask, then there must be something occluding the object, and that object's modal and amodal masks are candidates to be included in the dataset. Here we automatically include objects with $S(A_i) > 1.2\ S(M_i)$, *i.e.*, the area of its amodal mask is more than 1.2 times larger than its modal mask. Take the chair in Figure 3 as an example, we first generate its modal and amodal mask using Equations 1 and 2. Because the amodal mask of the chair is larger than its modal mask (it is occluded by the bed), we select the chair in the dataset. In this way, we have an automatic method to collect ground truth amodal masks for occluded objects in real images without any manual guessing.

Manual Selection. However, not all generated modal and amodal masks are of very good quality as the 3D meshes in MatterPort3D can be incomplete or noisy sometimes. We thus apply a manual selection stage, where human annotators inspect and select the pairs with good-quality modal and amodal masks. Bad quality examples due to problems of MatterPort3D are categorized (*e.g.,* the modal mask does not contain all visible parts of the object, or the amodal mask is noisy / incomplete) and shown to the human annotators. The manual selection is described in full detail in [50].

4. Architectures for Amodal Prediction

Given a single image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ and its corresponding modal (*visible*) mask $M_i \in \mathbb{R}^{H\times W}$ for the i-th object, our goal is to predict the amodal (*full*) mask for the object, $A_i \in \mathbb{R}^{H\times W}$. Specifically, we explore two architecture variants:

- A two-stage architecture, as shown in Figure 5, consisting of an **occluder predictor** to first estimate the occluder mask, followed by an **amodal predictor** to infer the amodal mask, given the modal mask, estimated occluder, and image.
- A one-stage architecture, as shown in Figure 6, that exploits the strong representation power of the pre-trained stable diffusion model, and adapts it to infer the amodal mask from the given image and modal mask.

4.1. Two-Stage Architecture – OccAmodal

Occluder Predictor. Occlusion in an image occurs when an object hides a part of another object, referred to as occluder and occludee respectively. For amodal completion, having the occluder's mask can largely simplify the task, as it provides information on which parts of one specified object should be completed [31, 51]. In existing works [31, 51], the occluder mask is often considered as a prior, and is directly fed into the model as input. One obvious limitation, however, is that the occluder mask can be unavailable at inference time. For example, in large-scale datasets, *e.g.* COCO [28] or LVIS [12], not all objects in an image are annotated, resulting in a failure of amodal completion in existing works [31, 51], *i.e.,* they cannot expand the modal mask at all if the segmentation of occluder mask is not annotated and provided. Here, instead of relying on an a-prior occluder mask, we consider a two-stage architecture, that first infers the occluder mask from the given image and the target object's modal mask, and then generates an amodal completion with the occluder mask as guidance. Specifically, as shown in Figure 5 (left), the occluder predictor takes the original image and the object's modal mask as input, to the ResNet and U-Net encoder respectively, and is then concatenated and upsampled to generate the prediction of the occluder mask for the object, $F_i = \Psi_{\{\text{OCP}\}}(\mathcal{I}, M_i)$, where $F_i \in \mathbb{R}^{H\times W}$ denotes the binary mask of occluder, that can be completely empty (no occluder), or with the union of all occluders.

Amodal Predictor. Given the mask of the predicted occluder, we compute the occlusion boundary (B_i), between the modal mask and occluder mask. We then feed the input

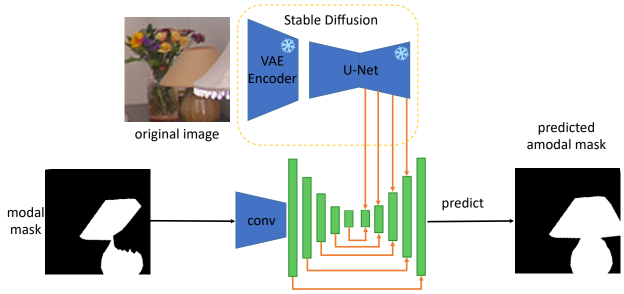

Figure 6. **One-Stage Architecture (SDAmodal) for Amodal Prediction.** The image is fed into a pre-trained Stable Diffusion model to get multi-scale representations containing occlusion information. The image and modal mask features are concatenated and forwarded to multiple decoding layers for amodal prediction. The Stable Diffusion model is frozen during training.

image, object's modal mask, and occlusion boundary to an amodal predictor, as shown in Figure 5 (right), similar to existing work [31]. In detail, both the input modal mask (M_i) and occlusion boundary (B_i) are concatenated, and input to a U-Net for encoding and decoding, with skip connections. Additionally, we also encode the input image with a ResNet, and inject it into the U-Net's bottleneck layer, providing visual conditioning for amodal completion. We denote the procedure as : $A_i = \Psi_{\{\text{AMP}\}}(\mathcal{I}, M_i, F_i)$

4.2. One-stage Architecture – SDAmodal

In recent literature, generative models based on diffusion have demonstrated the ability to generate photorealistic images, with seemingly correct geometry, object semantics and shapes. Here, we investigate the possibility of exploiting the visual features in diffusion models for amodal completion – after all, the task is mainly about understanding the shape of certain objects. Specifically, as shown in Figure 6, we feed the image into a pre-trained Stable Diffusion model, and add noise onto the latent features after the VAE encoder. We extract the multi-scale features from the decoding layers of the U-Net at time step 181 following the investigation about occlusion in [49]. Then we concatenate the Stable Diffusion features with multi-scale features of the modal mask, and forward them to multiple decoding layers to generate the amodal mask prediction, the procedure can be denoted as : $A_i = \Psi_{\{\text{SD}\}}(\mathcal{I}, M_i)$.

4.3. Training

Training the first stage of OccAmodal requires ground truth occluder masks, while training both the second stage of OccAmodal and the SDAmodal requires the ground truth amodal masks. Both COCOA and our MP3D-Amodal provide ground truth amodal masks while only COCOA provides ground truth occluder masks.

OccAmodal. For training of the occluder predictor, the occlusion relationships annotated in COCOA [53] are used to obtain the ground truth occluder mask, and then the pixel-level prediction of occluder mask is trained via cross-entropy loss. For training of the amodal predictor, the amodal mask prediction is supervised by the ground truth amodal mask (provided by COCOA or MP3D-Amodal) via an Uncertainty Weighted Segmentation Loss as mentioned in [31].

SDAmodal. For training of the Stable Diffusion based architecture, the amodal mask prediction is supervised by the ground truth amodal mask (provided by COCOA or MP3D-Amodal) via a cross-entropy loss.

5. Experiments

5.1. Experimental Details

Datasets and Implementation Details. We employ both COCOA [53] and our collected MP3D-Amodal (Section 3) for training and evaluating our models. To ensure a fair comparison, we use the same training setting as in [31, 51], which employs SGD with momentum, sets the learning rate to be $1e^{-3}$, and trains the model for 56K iterations with a batch size of 32. Models are trained on A6000 / A40 GPUs. More training details are given in the [50].

Baselines. We compare with two existing amodal completion models [31, 51], of which [31] is the latest state-of-the-art method for amodal completion. [51] has both one-stage and two-stage architectures, which we denote as Deocclusion (Single Stage) and Deocclusion (Two Stage). The default architecture in [51] is Deocclusion (Two Stage), while Deocclusion (Single Stage) uses a ResNet to encode the input image and concatenate it with the features in the U-Net decoder (similar to Figure 5 left). Additionally, in [50], we compare with recent amodal instance segmentation methods such as VRSP [45], A3D [25], AISformer [40], C2F-Seg [10] and GIN [26].

Evaluation. Following [31, 51], we compute mIOU between the ground truth and predicted amodal mask. Additionally, mIOU-inv is also used, which refers to the mIOU for only the occluded regions.

5.2. Ablation Study of Different Architectures

In Table 3, we ablate the importance of the occluder predictor and the number of skip connections for the OccAmodal architecture. As can be seen, the occluder mask is crucial for amodal mask prediction. This is eveident from the results of Setting $\mathbb{A}$, achieving only 69.9 mIOU on COCOA. In comparison, when the predicted occluder mask is incorporated, amodal completion can be boosted to 88.4 mIOU (Setting $\mathbb{B}$) on COCOA, and the performance is further boosted when we include a skip connection for the final layer of the U-Net (Setting $\mathbb{C}$). In the architecture of [31] there are only 4 skip connections and we are adding the fifth. In Table 4, we

ID	Occluder Predictor	Final Skip Connection	COCOA	
			mIOU	mIOU-inv
$\mathbb{A}$			69.9	0.006
$\mathbb{B}$	✓		88.4	64.4
$\mathbb{C}$	✓	✓	**89.4**	**66.2**

Table 3. **Ablation Study of OccAmodal.** Setting $\mathbb{A}$ is the setting of ASBU [31]. All models are trained on COCOA.

ID	SD Backbone	Multi-Scale SD Feature	Final Skip Connection	COCOA	
				mIOU	mIOU-inv
$\mathbb{A}$				88.0	63.8
$\mathbb{B}$	✓			89.4	69.2
$\mathbb{C}$	✓	✓		89.6	69.8
$\mathbb{D}$			✓	89.2	66.4
$\mathbb{E}$	✓		✓	90.5	71.1
$\mathbb{F}$	✓	✓	✓	**90.7**	**71.6**

Table 4. **Ablation Study of SDAmodal.** Setting $\mathbb{A}$ is Deocclusion (Single Stage) in [51]. All models are trained on COCOA.

ID	Comments	Occluder Mask Provided	COCOA		MP3D-Amodal	
			mIOU	mIOU-inv	mIOU	mIOU-inv
$\mathbb{A}$	Deocclusion(Two Stage) [51]	✓	88.2	65.3	-	-
$\mathbb{B}$	ASBU [31](reproduced)	✓	88.9	65.3	-	-
$\mathbb{C}$	ASBU [31](reported)	✓	89.9	-	-	-
$\mathbb{D}$	Deocclusion (Two Stage) [51]		69.9	0.006	64.4	0.004
$\mathbb{E}$	ASBU [31]		69.9	0.006	64.4	0.004
$\mathbb{F}$	Deocclusion (Single Stage) [51]		88.0	63.8	72.4	28.0
$\mathbb{G}$	OccAmodal		89.4	66.2	72.9	27.5
$\mathbb{H}$	SDAmodal		**90.7**	**71.6**	**76.4**	**38.5**

Table 5. **Comparison with State-of-the-Art Amodal Completion Methods**. Our SDAmodal model achieves the new state-of-the-art performance for amodal completion over a larger variety of categories. All models are trained on COCOA, and evaluated on both COCOA and MP3D-Amodal.

ablate variations on the SDAmodal architecture. Replacing the original ResNet image encoder with the Stable Diffusion backbone brings a significant boost (+1.4/+5.4 in terms of mIOU and mIOU-inv for Setting $\mathbb{A}$ to $\mathbb{B}$, +1.3/+4.7 in terms of mIOU and mIOU-inv for Setting $\mathbb{D}$ to $\mathbb{E}$). If multiple layers of Stable Diffusion features at different resolutions are fed into the model (as shown in Figure 6) the performance is higher than if only a single layer feature is used (the second layer of the Stable Diffusion U-Net as in [49]) (comparing Setting $\mathbb{B}$/$\mathbb{C}$ and $\mathbb{E}$/$\mathbb{F}$). The performance can also be improved by adding a final layer skip connection for the U-Net (comparing Setting $\mathbb{A}$/$\mathbb{D}$, $\mathbb{B}$/$\mathbb{E}$ and $\mathbb{C}$/$\mathbb{F}$). According to [49], the features of other pre-trained models such as DINO [2, 33] and CLIP [15, 36] perform worse than Stable Diffusion features on "occlusion" task. We have further trained our model using DINO and CLIP features. The results are given in the ArXiv version of this paper [50], validating the superiority of Stable Diffusion features.

5.3. Comparison with State-of-the-Art

We compare our method with previous amodal completion state-of-the-art methods, Deocclusion [51] and ASBU [31], on both COCOA and MP3D-Amodal. Note that, ASBU [31] and Deocclusion (Two Stage) [51] require the occluder masks provided, while in both of the architectures we propose, the occluder masks are not necessary. The comparisons are given in Table 5. We can make the following observations: (1) SDAmodal outperforms the previous state-of-the-art methods (Setting $\mathbb{C}$ and $\mathbb{H}$) even if the occluder mask is not provided for SDAmodal, but is for previous methods; (2) When the occluder mask is not provided, previous methods Deocclusion (Two Stage) [51] and ASBU [31] cannot expand the modal mask of the object and achieve poor performance for amodal completion (Setting $\mathbb{D}$ and $\mathbb{E}$). In comparison, OccAmodal (Setting $\mathbb{G}$), where the occluder mask is generated by our occluder predictor, has a high performance, demonstrating the effectiveness of the occluder mask prediction module. (3) Even though SDAmodal is only trained on COCOA, the Stable Diffusion backbone efficiently boosts the performance not only on COCOA, but also *zero-shot generalized* to MP3D-Amodal which contains objects from different domains and categories (compare Settings $\mathbb{F}$ and $\mathbb{H}$ where the difference is +4.0/+10.5 in terms of mIOU and mIOU-inv). We have also tested our models on the COCOA-cls benchmark [9] for amodal instance segmentation on 80 COCO categories and show superior performance. Please refer to ArXiv version of this paper [50] for more details.

5.4. Effectiveness of Different Training Data

Table 6 shows the effectiveness of training with extra data from our MP3D-Amodal training split. Both OccAmodal and SDAmodal improve performance on MP3D-Amodal when they are also trained with MP3D-Amodal and there is no performance deterioration on COCOA.

5.5. Qualitative Results

In Figure 7, we show a qualitative comparison of different amodal completion methods on both the COCOA and MP3D-

ID	Architecture	COCOA	MP3D-Amodal	COCOA		MP3D-Amodal	
				mIOU	mIOU-inv	mIOU	mIOU-inv
$\mathbb{A}$	OccAmodal	✓		**89.4**	66.2	72.9	27.5
$\mathbb{B}$	OccAmodal	✓	✓	**89.4**	**66.4**	**73.8**	**29.6**
$\mathbb{C}$	SDAmodal	✓		**90.7**	**71.6**	76.4	38.5
$\mathbb{D}$	SDAmodal	✓	✓	**90.7**	**71.6**	**81.8**	**53.7**

Table 6. **Effectiveness of Different Training Data**. The performance of both models are boosted on MP3D-Amodal if extra training data from MP3D-Amodal is used.

Figure 7. **Qualitative Comparison on COCOA and MP3D-Amodal**. COCOA: Rows 1, 2 and 3; MP3D-Amodal: Rows 4 and 5. Please see the text for more discussion. More qualitative examples are provided in [50].

Amodal datasets. We observe that ASBU [31] faces limitations in expanding the modal mask when the occluder mask is not provided (Column 3). Deocclusion (Single Stage) can partially complete the amodal mask when the occluder mask is not available but the prediction quality is not good (Column 4). In contrast, our models, especially SDAmodal, can handle the situation where the occluder mask is not provided and significantly improve the accuracy of amodal mask predictions (Columns 5 and 6), even when the object to complete is from a different domain (Rows 4 and 5) when only trained on COCOA. More qualitative results are in [50].

6. Conclusion and Extensions

By utilising real 3D data, we have proposed an automatic pipeline to generate ground truth amodal masks for occluded objects in real images, and used this to create a new amodal segmentation evaluation benckmark for a large variety of instances. The pipeline has been applied to the MatterPort3D dataset, but can be applied to other suitable datasets such as ScanNet [6] that have real images together with the 3D structure for objects in the scene. Furthermore, we have developed two new state-of-the-art methods for amodal completion *in the wild* – *i.e.*, capable of handling situations where the occluder is unknown or undefined, and for a wide variety of object classes. The models, with a lightweight occluder predictor and Stable Diffusion representations, achieve superior performance on different domains and object categories.

Acknowledgements. This research is supported by EPSRC Programme Grant VisualAI EP/T028572/1, a Royal Society Research Professorship RP\R1\191132, an AWS credit funding, a China Oxford Scholarship and ERC-CoG UNION 101001212. We also thank Rajan from Elancer and his team, for their huge assistance with annotation.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: high quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 43(5):1483–1498, 2019. 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 7

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 2, 3, 4

[4] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 4974–4983, 2019. 2

[5] 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 (CVPR)*, pages 1290–1299, 2022. 1, 2

[6] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017. 8

[7] Helisa Dhamo, Nassir Navab, and Federico Tombari. Object-driven multi-layer scene decomposition from a single image. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[8] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. SeGAN: Segmenting and generating the invisible. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6144–6153, 2018. 2

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

[10] Jianxiong Gao, Xuelin Qian, Yikai Wang, Tianjun Xiao, Tong He, Zheng Zhang, and Yanwei Fu. Coarse-to-fine amodal segmentation with shape prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1262–1271, 2023. 2, 6

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361. IEEE, 2012. 2

[12] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 5356–5364, 2019. 5

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 2961–2969, 2017. 1, 2

[14] Cheng-Yen Hsieh, Tarasha Khurana, Achal Dave, and Deva Ramanan. Tracking any object amodally. *arXiv preprint arXiv:2312.12433*, 2023. 1

[15] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 7

[16] Yusuke Inagaki, Ryosuke Araki, Takayoshi Yamashita, and Hironobu Fujiyoshi. Detecting layered structures of partially occluded objects for bin picking. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5786–5791. IEEE, 2019. 2

[17] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018. 1

[18] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1966–1974, 2015. 1

[19] 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 (CVPR)*, pages 4019–4028, 2021. 2

[20] Philip J Kellman and Christine M Massey. Perceptual learning, cognition, and expertise. In *Psychology of learning and motivation*, pages 117–165. Elsevier, 2013. 2

[21] Sung-Kyun Kim and Maxim Likhachev. Planning for grasp selection of partially occluded objects. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3971–3978. IEEE, 2016. 2

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 3

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 2

[24] Ke Li and Jitendra Malik. Amodal instance segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 677–693. Springer, 2016. 2

[25] Zhixuan Li, Weining Ye, Tingting Jiang, and Tiejun Huang. 2d amodal instance segmentation guided by 3d shape prior. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 165–181. Springer, 2022. 2, 6

[26] Zhixuan Li, Weining Ye, Tingting Jiang, and Tiejun Huang. Gin: Generative invariant shape prior for amodal instance

segmentation. *IEEE Transactions on Multimedia (MM)*, 2023. 2, 6

[27] Zhixuan Li, Weining Ye, Juan Terven, Zachary Bennett, Ying Zheng, Tingting Jiang, and Tiejun Huang. Muva: A new large-scale benchmark for multi-view amodal instance segmentation in the shopping scenario. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23504–23513, 2023. 2

[28] 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 European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 2, 5

[29] Huan Ling, David Acuna, Karsten Kreis, Seung Wook Kim, and Sanja Fidler. Variational amodal object completion. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:16246–16257, 2020. 1, 3

[30] Rohit Mohan and Abhinav Valada. Amodal panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21023–21032, 2022. 2

[31] Khoi Nguyen and Sinisa Todorovic. A weakly supervised amodal segmenter with boundary uncertainty estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7396–7405, 2021. 1, 2, 3, 5, 6, 7, 8

[32] Pedro O O Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. *Advances in neural information processing systems (NeurIPS)*, 28, 2015. 2

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 7

[34] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. *arXiv preprint arXiv:2401.14398*, 2024. 1

[35] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3014–3023, 2019. 2

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[37] N Dinesh Reddy, Robert Tamburo, and Srinivasa G Narasimhan. Walt: Watch and learn 2d amodal representation from time-lapse imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9356–9366, 2022. 2

[38] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *Proceedings of the European Conference on Computer Vision (ECCV)*, 7576:746–760, 2012. 1

[39] Yihong Sun, Adam Kortylewski, and Alan Yuille. Amodal segmentation through out-of-task and out-of-distribution generalization with a bayesian model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1215–1224, 2022. 2

[40] Minh Tran, Khoa Vo, Kashu Yamazaki, Arthur Fernandes, Michael Kidd, and Ngan Le. Aisformer: Amodal instance segmentation with transformer. *British Machine Vision Conference (BMVC)*, 2022. 2, 6

[41] Basile Van Hoorick, Pavel Tokmakov, Simon Stent, Jie Li, and Carl Vondrick. Tracking through containers and occluders in the wild. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023. 1

[42] Jacob Varley, Chad DeChant, Adam Richardson, Joaquín Ruales, and Peter Allen. Shape completion enabled robotic grasping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2442–2447, 2017. 2

[43] Ruihai Wu, Kai Cheng, Yan Zhao, Chuanruo Ning, Guanqi Zhan, and Hao Dong. Learning environment-aware affordance for 3d articulated object manipulation under occlusions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[44] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. MagicPony: Learning articulated 3d animals in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023. 1

[45] 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 (AAAI)*, pages 2995–3003, 2021. 2, 6

[46] Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting moving objects via an object-centric layered representation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[47] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Amodal completion via progressive mixed context diffusion. *arXiv preprint arXiv:2312.15540*, 2023. 1

[48] Guanqi Zhan, Weidi Xie, and Andrew Zisserman. A tri-layer plugin to improve occluded detection. *British Machine Vision Conference (BMVC)*, 2022. 1

[49] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. What does stable diffusion know about the 3d scene? In *arXiv:2310.06836*, 2023. 6, 7

[50] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. Amodal ground truth and completion in the wild. *arXiv preprint arXiv:2312.17247*, 2023. 3, 5, 6, 7, 8

[51] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3784–3792, 2020. 1, 2, 3, 5, 6, 7

[52] Chuanxia Zheng, Duy-Son Dao, Guoxian Song, Tat-Jen Cham, and Jianfei Cai. Visiting the invisible: Layer-by-layer completed scene decomposition. *International Journal of Computer Vision (IJCV)*, 129:3195–3215, 2021. 2

[53] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1464–1472, 2017. 2, 6

[54] Wei Zhuo, Mathieu Salzmann, Xuming He, and Miaomiao Liu. Indoor scene parsing with instance segmentation, semantic labeling and support relationship inference. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6269–6277, 2017. 1

[55] Chuhang Zou and Derek Hoiem. Silhouette guided point cloud reconstruction beyond occlusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020. 1