Amodal Scene Analysis via Holistic Occlusion Relation Inference and Generative Mask Completion

Bowen Zhang[1], Qing Liu[2], Jianming Zhang[2], Yilin Wang[2], Liyang Liu[1], Zhe Lin[2], Yifan Liu[1†]

[1]Australian Institute for Machine Learning, The University of Adelaide
[2] Adobe Research
{b.zhang,akide.liu,yifan.liu04}@adelaide.edu.au, {qingl,jianmzha,yilwang,zlin}@adobe.com

Abstract

Amodal scene analysis entails interpreting the occlusion relationship among scene elements and inferring the possible shapes of the invisible parts. Existing methods typically frame this task as an extended instance segmentation or a pair-wise object de-occlusion problem. In this work, we propose a new framework, which comprises a Holistic Occlusion Relation Inference (HORI) module followed by an instance-level Generative Mask Completion (GMC) module. Unlike previous approaches, which rely on mask completion results for occlusion reasoning, our HORI module directly predicts an occlusion relation matrix in a single pass. This approach is much more efficient than the pair-wise de-occlusion process and it naturally handles mutual occlusion, a common but often neglected situation. Moreover, we formulate the mask completion task as a generative process and use a diffusion-based GMC module for instance-level mask completion. This improves mask completion quality and provides multiple plausible solutions. We further introduce a large-scale amodal segmentation dataset which consists of high-quality human annotations for amodal masks and occlusion relations, including mutual occlusions. Experiments on the newly proposed dataset and two public benchmarks demonstrate the advantages of our method on both efficient occlusion reasoning and plausible amodal mask completion. code public available at https://github.com/zbwxp/Amodal-AAAI.

Introduction

Humans can naturally perceive the occlusion relationship among multiple scene elements and infer the possible shapes for the invisible parts, and this ability is known as amodal perception (Nanay 2018; Mohan and Valada 2022; Zhu et al. 2017). In computer vision, amodal scene analysis has been proposed to match the human intelligence (Li and Malik 2016), and this task can provide useful information for many real-world applications. For example, occlusion reasoning can facilitate risk assessment in AI systems, where existing methods are data-driven and may become less reliable when applied on heavily occluded objects (Zhu et al. 2019; Kortylewski et al. 2020). Amodal scene analysis can also benefit image editing tasks. By decomposing the 2D scene into layer-wise representations based on the predicted occlusion relationship and amodal shapes, users can easily move things around and generate new RGB content (Zhan et al. 2020a; Zheng et al. 2021).

Many existing works try to solve amodal problems by extending traditional instance segmentation methods, such as Mask R-CNN (He et al. 2017) and DETR (Carion et al. 2020). In those methods, heads for amodal mask prediction are added to the instance segmentation model (Follmann et al. 2019; Qi et al. 2019; Xiao et al. 2021; Tran et al. 2022). These methods treat amodal segmentation as an object recognition task, and entangle instance segmentation and amodal mask completion in a single framework. Consequently, the problem becomes too challenging and satisfying results can hardly be obtained. In addition, these methods only focus on amodal mask prediction and do not explicitly address the occlusion reasoning problem, which is an essential part for amodal scene analysis. They also often use mean Average Precision (mAP) as the main evaluation metric, which cannot reflect the mask quality or fidelity in many cases. Furthermore, since these methods are often trained for a set of predefined object classes, their application is strictly constrained by the training data.

Another approach decouples amodal analysis from a perspective on object recognition. Taking instance masks of visible regions as input, these methods interpret the occlusion relationship among the objects of interest and perform amodal mask completion in a separate step (Yan et al. 2019a; Ling et al. 2020; Zhan et al. 2020b; Nguyen and Todorovic 2021). These approaches focus more on amodal analysis and can generally achieve better results for occlusion reasoning and mask completion. However, current approaches in this direction typically rely on a time-consuming process of pair-wise object de-occlusion. The occlusion order reasoning is based on mask completion results, making the overall performance unstable due to the difficulty of the amodal segmentation task. In addition, many prior works utilize a deterministic method to solve amodal mask completion, which is limited in its ability to capture multiple potential and plausible shapes that may exist under occlusion.

In this work, we propose a new amodal scene analysis framework based on the second approach, where we decouple the problem from object recognition. Since many existing works (Cheng et al. 2022; Jain et al. 2022; Li et al. 2022a) on instance and panoptic segmentation tasks already achieved impressive performance, in this work we assume

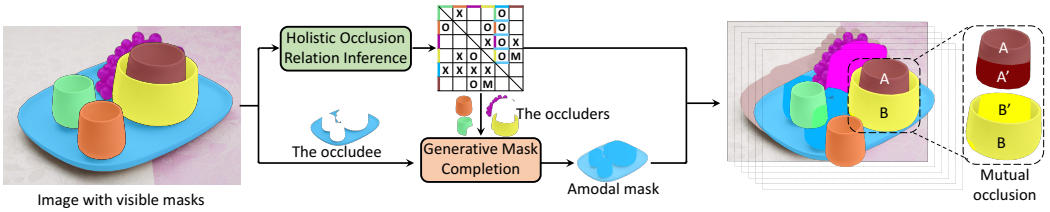


Figure 1: Overview of the proposed framework. Given an image with visible object masks, we propose the Holistic Occlusion Relation Inference (HORI) module to infer the occlusion relationship matrix for all objects in a single pass. The colored edges of the rows and columns in the matrix match the corresponding colored object mask on the left. 'O' and 'X' in the matrix indicate whether an object is occluding or occluded by another object, while 'M' indicates mutual occlusion. For instance, region $\mathbf{A}'$ is occluded by $\mathbf{B}$, and $\mathbf{B}'$ is occluded by $\mathbf{A}$. The occlusion relation matrix, along with the image and its visible masks, is then fed to our Generative Mask Completion (GMC) module to generate the amodal masks at the instance level.

instance segmentation masks are available for objects of interest in the scene. As shown in Fig. 1, we first introduce a Holistic Occlusion Relation Inference (HORI) module which achieves single-pass occlusion reasoning by directly predicting an occlusion relationship matrix. We then present a Generative Mask Completion (GMC) module which performs instance-level amodal completion through a diffusion-based sampling process and enables multiple plausible outputs. More specifically, the HORI module adapts the architecture of Mask2former (Cheng et al. 2022) and takes instance masks as additional inputs. It leverages the attention mechanism of Mask2former and attends to the visible region of each instance directly to learn a dense ordering relationship. By outputting an occlusion relationship matrix, the module is much more efficient than the previous pair-wise de-occlusion methods (Yan et al. 2019a; Ling et al. 2020; Zhan et al. 2020b; Nguyen and Todorovic 2021) and can naturally handle mutual occlusion, which is a common but often neglected situation. Then, given the instance masks and ordering relationship for objects in the scene, we apply a GMC module to perform instance-level amodal mask completion. The GMC module adapts a diffusion-based sampling process and is conditioned on the visible mask to infer multiple potential shapes of an amodal mask. Compared with previous deterministic methods, the GMC module also produces amodal masks with higher fidelity.

In our experiments, we evaluate our model on two popular amodal benchmarks, COCOA (Zhu et al. 2017) and KINS (Qi et al. 2019), where we demonstrate that our HORI module achieves new state-of-the-art results for Ordering Accuracy (O-Acc) with highly efficient single-pass inference. Additionally, our GMC module outperforms existing methods not only on mean Intersection-over-Union (mIoU), but also on more advanced metrics that evaluate the fidelity of the predicted shapes for amodal mask completion. To facilitate a more comprehensive model evaluation for amodal scene analysis, especially for the case of mutual occlusion, we further introduce a large-scale amodal segmentation dataset, Amodal Scene in the Wild (ASW). The evaluation set of ASW will be released with the paper, which consists of $2,000$ images of diverse scenes and $14,969$ high-quality amodal masks with occlusion ordering. Among the $13,240$ occlusion relationships revealed in the annotation, $2,515$ are mutual. We demonstrate our proposed method can achieve accurate mutual occlusion prediction both quantitatively and qualitatively on the ASW dataset.

We summarize our main contributions as follows:

- We propose a new framework to solve amodal scene analysis by two modules. The Holistic Occlusion Relation Inference (HORI) module interprets the occlusion relationship among multiple scene elements in a single pass, while the Generative Mask Completion (GMC) module predicts diverse high-fidelity amodal shapes by formulating the task as a generative sampling process.
- We investigate mutual occlusion, a common situation in real-world scenes but largely overlooked by previous methods and benchmarks. To this end, we introduce a new dataset that consists of diverse scenes and high-quality annotations for amodal masks and occlusion relations, including $2,515$ mutual occlusion cases.

Related Work

Occlusion Reasoning. A number of works have studied occlusion reasoning as a multi-view problem (Kang, Szeliski, and Chai 2001; Yamaguchi, McAllester, and Urtasun 2014; Gilroy, Jones, and Glavin 2019) while inferring occlusion relationship from a single image is more challenging (Hoiem et al. 2007; Hsiao and Hebert 2014; Jiang et al. 2020). Earlier works have explored prior and template-based methods (Tighe, Niethammer, and Lazebnik 2014; Wu, Tenenbaum, and Kohli 2017). Another line of works uses closed contour to express the object, and the orientation of each contour pixel to describe the order relationship (Ren, Fowlkes, and Malik 2006; Wang and Yuille 2016; Lu et al. 2019). In the domain of amodal analysis, the ordering recovery method introduced in (Zhan et al. 2020b) works by learning the relationship between pairwise synthetic data and it has inspired several later works, such as (Nguyen and Todorovic 2021; Yan et al. 2019a).

Amodal Segmentation and Mask Completiong. Amodal segmentation aims at detecting objects and recovering their complete shapes. Many existing works inherit the network architecture from popular instance segmentation methods and apply amodal mask supervision on top of the training (Zhu et al. 2017; Qi et al. 2019; Sun, Kortylewski, and Yuille 2022; Xiao et al. 2021; Mohan and Valada 2022). These methods usually apply multiple layers of convolu-

tions and deformable convolutions to strengthen the model's ability to infer the invisible region. Although they perform well on some simple rigid objects, they are more likely to fail on objects with irregular or elongated shapes (e.g. table legs and human arms). Another line of work focuses on amodal mask completion, where instance segmentation masks are provided as input. This approach simplifies the problem by decoupling amodal analysis from object recognition and thus can achieve better results. Besides early unsupervised contour completion methods that are constrained on toy examples (Kimia, Frankel, and Popescu 2003; Silberman et al. 2014), 3D templates and synthetic data have been used broadly (Kar et al. 2015; Ehsani, Mottaghi, and Farhadi 2018; Yan et al. 2019b). (Zhan et al. 2020b) proposes a self-supervised scene de-occlusion method by learning amodal masks from 2D synthetic occlusions, which is followed by (Nguyen and Todorovic 2021; Yan et al. 2019a), while (Ling et al. 2020) learns from synthetic data similarly but develops a variational generative framework for the task.

Diffusion Models. Diffusion probabilistic models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020) employ a forward Markov chain to diffuse the data to noise and learn the reversal of such a diffusion process. Conditional diffusion models encode additional information (e.g., semantic layout) into the generation process and improve largely the generation performance, inspiring a variety of tasks including image generation (Jolicoeur-Martineau et al. 2020; Dhariwal and Nichol 2021; Vahdat, Kreis, and Kautz 2021; Rombach et al. 2022; Ho et al. 2022), image editing (Nichol et al. 2021; Saharia et al. 2022a; Kawar et al. 2022; Couairon et al. 2022; Zeng et al. 2022), super-resolution (Li et al. 2022b; Saharia et al. 2022b), etc. In this work, we use diffusion model to complete binary masks instead of generating RGB values, which has rarely been explored.

Methods

The task of amodal segmentation involves generating the amodal mask $\mathbf{M}_{amodal}$ for an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and its corresponding instance masks $\mathbf{M}_{inst} \in \mathbb{R}^{H\times W\times N}$, where N represents the number of objects present in the image. The instance masks provide information about the visible areas of the objects, while the amodal masks provide the complete shapes, including the occluded regions.

To achieve this, our approach involves first inferring the occlusion relationships between all the objects using holistic occlusion relation inference. This provides a holistic understanding of how each object occludes other objects in the image. This occlusion relationship is then used as input for the generative mask completion module, which generates the complete amodal mask $\mathbf{M}_{amodal}$ for each object. It should be noted that the inference of instance masks $\mathbf{M}_{inst}$ is typically done using instance or entity segmentation methods, which are not the focus of this paper. Instead, we focus on the generation of amodal masks, which is a critical task in many computer vision applications such as object tracking, scene understanding, and robotic perception.

Holistic Occlusion Relation Inference Module

To enable concurrent ordering, an efficient method is required to encode arbitrary numbers of binary masks into a model structure for further processing. As illustrated in Fig. 2, our framework first passes the image through a Resnet50 backbone and a deformable-DETR decoder to produce feature maps of different levels with rich semantic information. The binary masks $\mathbf{M}_{inst}$ indicating the object's visible parts are provided, and the number of masks N can vary for different images. We dynamically duplicate the initial token N times to match the number of visible masks, and these tokens are then passed through the Occlusion Reasoning Block (ORB).

Each ORB block requires three inputs: the feature map $\mathbf{F} \in \mathbb{R}^{h\times w\times C}$ where C representing the number of channels of the feature map, the visible binary masks $\mathbf{M}_{inst}$, and the tokens $\mathbf{T} \in \mathbb{R}^{N\times C}$ to be matched with the visible masks. These three inputs are first passed through a masked multi-head cross-attention (MMHCA) module, where $\mathbf{T}$ serves as queries, $\mathbf{F}$ as keys and values, and $\mathbf{M}_{inst}$ as the attention mask. The computation is as follows with l indicating the layer index.

$$\begin{aligned} &\mathbf{Q}_l = f_Q(\mathbf{T}_l), \mathbf{K}_l = f_K(\mathbf{F}_l), \mathbf{V}_l = f_V(\mathbf{F}_l). \\ &\mathcal{M} = \begin{cases} 0 & \text{if } \mathbf{M}_{inst} = 1 \\ -\infty & \text{otherwise} \end{cases} \\ &\mathbf{T}_{l+1} = \text{softmax}\left(\mathcal{M} + \mathbf{Q}_l\mathbf{K}_l^T\right)\mathbf{V}_l + \mathbf{T}_l \end{aligned} \tag{1}$$

This masked attention will gather the information in $\mathbf{F}_l$ within the $\mathbf{M}_{inst}$ and update the $\mathbf{T}_l$. During the inference, the first $\mathbf{T}_0$ is generated by repeating a random initialized learnable token embedding for N times. After the first ORB block, the originally identical tokens will become diverse and gradually correspond to each mask region in $\mathbf{M}_{inst}$. We use this mechanism to encode the spatial information of $\mathbf{M}_{inst}$ into the model. The output $\mathbf{T}_{l+1}$ is then carried out by a self-attention module and a feed-forward network (FFN) as in a regular transformer decoder.

Each ORB block produces three outputs: an occlusion relation matrix, N occlusion ratio predictions, and N binary mask predictions. The occlusion relation matrix is generated from the attention map of the multi-head self-attention (MHSA) process on $\mathbf{T}$. The attention map naturally encodes the ordering relationships between all elements in $\mathbf{T}$, and we use it to generate the occlusion relation matrix. To introduce more non-linearity, we attach an additional MLP module to the attention map. The produced occlusion relation matrix is then used for ordering recovery. The occlusion ratio prediction and the binary mask predictions of the ORB block follow a similar structure to DETR. First, the $\mathbf{T}$ tokens are attached to a fully-connected layer that predicts the occlusion ratio of the current token. Then, the output is attached to another MLP module that generates the kernel for a 1×1 convolution operation. The kernel is then applied to the $1/4$ resolution feature map to generate a binary mask prediction, representing the current token's amodal prediction. Those outputs are then supervised with their corresponding amodal masks and occlusion ratios sequentially where nine ORB

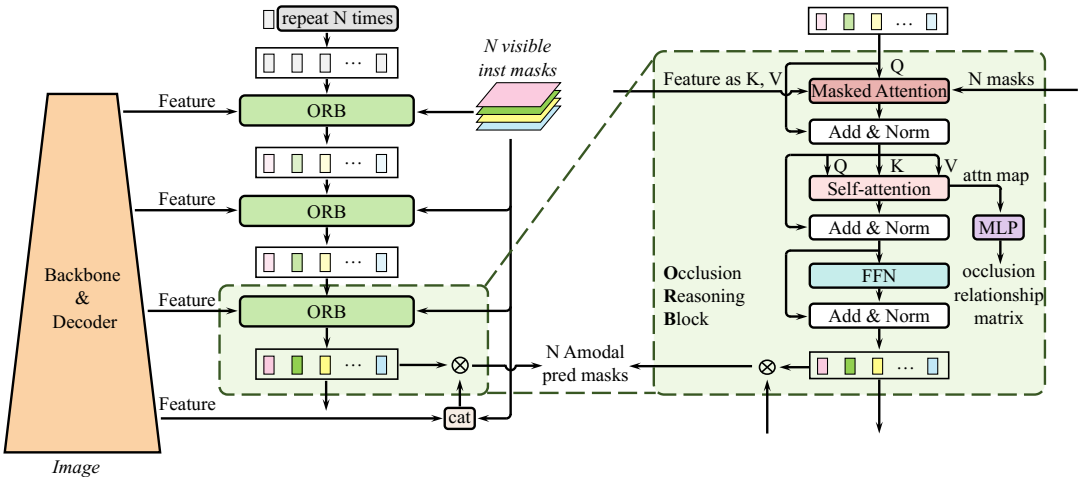


Figure 2: The overall structure of the Holistic Occlusion Relation Inference (HORI) module. The HORI module combines a Resnet50 backbone and a deformable DETR decoder to extract variable-resolution features from input images. Then the core component, the Occlusion Reasoning Block (ORB), employs feature maps as keys and values and employs N visible instance masks for masked attention during the learning of token embeddings, which are later used for occlusion reasoning. Importantly, the HORI module's role is to incorporate visible instances and amodal masks into the attention process through the ORB. It does NOT aim to predict amodal masks but rather enhances occlusion reasoning capabilities.

blocks are attached to feature maps of three resolutions sequentially, with each ORB having its corresponding supervision.

Training Supervisions. The corresponding losses for the ORB block are as follows:

$$\begin{aligned} \mathcal{L}_{mask} &= \mathcal{L}_{CE} + \mathcal{L}_{dice} \\ \mathcal{L}_{matrix} &= \mathcal{L}_{CE} + \mathcal{L}_{dice} \\ \mathcal{L}_{all} &= \lambda_1 \mathcal{L}_{mask} + \lambda_2 \mathcal{L}_{matrix} + \lambda_3 \mathcal{L}_{ratio} \end{aligned} \tag{2}$$

The first loss function is $\mathcal{L}_{mask}$, which is inherited from DETR for amodal mask supervision. This loss function is used to ensure that the generated amodal masks are accurate and match the ground truth masks. The target for $\mathcal{L}_{mask}$ is the ground truth amodal mask.

The second loss function is $\mathcal{L}_{matrix}$, which is used to supervise the occlusion relationships between objects. The target for $\mathcal{L}_{matrix}$ is a $N \times N$ binary matrix indicating the occlusion relationships between the N objects in the image. This loss function is designed to ensure that the model correctly infers occlusion relationships between objects in the image. Since the target can also be treated as a binary mask, we use the same loss as in $\mathcal{L}_{mask}$.

The third loss function is $\mathcal{L}_{ratio}$, which is used to supervise the occlusion ratio of each object in the image. The occlusion ratio $r \in [0, 1)$ indicates the amount of occlusion a given object is experiencing, with 0 indicating no occlusion and 1 indicating complete occlusion. This loss function ensures that the model learns to accurately predict the occlusion ratio of each object in the image. The loss is computed using mean squared error. The weights $\lambda 1, \lambda_2, \lambda_3$ are hyperparameters that control the relative importance of each loss function in the training process.

Noting that during the generation of the amodal prediction, the binary mask $\mathbf{M}_{inst}$ is concatenated to the feature map. This approach can sometimes result in instances without occlusion converging to a trivial solution, where the visible mask channel has a very high weight. However, we intentionally designed the structure in this way to reduce the burden on the model to handle instances without occlusion. In experiments, we found this strategy to be highly effective.

The ORB block is essential for producing accurate ordering recovery, with all pairs playing a crucial role. The masked attention mechanism encodes the visible parts and trains the model to interpolate the invisible parts of the amodal prediction, resulting in reasonable ordering relationships. The ORB structure facilitates the tokenization of binary masks into tokens that pair with feature maps. Supervision of the amodal predictions helps the model encode visible information accurately and interpret the invisible parts. Lastly, supervision on the order map enables the model to learn the ordering relationships within all the objects.

Handling Mutual Occlusion. Existing datasets assume that the occlusion relationship between objects is bipartite, meaning that one object can only be the 'occluder' or the 'occludee' in one object pair. However, during our experiments, we found that this assumption is often invalid, and there are instances where the occlusion relationship between two objects is mutual, typically happening when objects are interacting with each other. Examples of mutual occlusion relationships can be seen in Fig. 3. To avoid mutual occlusion annotations, previous datasets such as COCOA (shown in the first row of the figure) split instances into multiple parts. However, this approach can result in inconsistent annotations (see Fig. 3). To construct our ASW dataset (see Sec.), we purposefully allow mutual occlusion to make the amodal mask annotation consistent.

Previous methods for pairwise ordering, such as those proposed in (Zhan et al. 2020b; Nguyen and Todorovic 2021), rely on comparing the sizes of the intruding areas

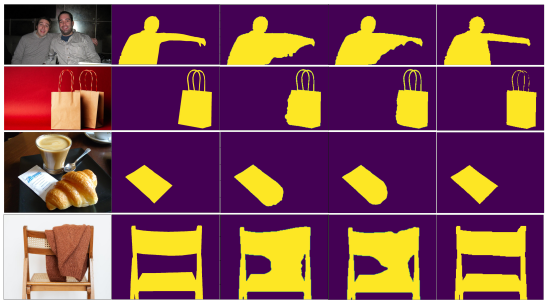
Figure 3: COCOA (1st row) splits instances into multiple parts (indexed by numbers) to avoid the need for mutual occlusion annotations, resulting in inconsistent instance and amodal masks. Our ASW dataset (2nd row) addresses this by annotating each instance with its complete shape directly.

of the predicted amodal mask of one object with the visible mask of the other object to make their ordering predictions. These methods were designed and trained on synthetic data with pairwise bipartite ordering relationships. It is non-trivial to generate synthetic data with mutual occlusion. As a result, these previous methods fail to adapt to our newly presented fully annotated ASW dataset. In contrast, our HORI module (Fig. 2) can easily handle the mutual occlusion prediction by simply having two output channels for the MLP in the ORB block to encode both 'occluder' and 'occludee' status, which can be true at the same time, for each object with respect to another one.

Generative Mask Completion Module

After successfully inferring the order relationships among the objects, the next step in our approach is to process the mask completion in order to obtain the final amodal mask predictions. Previous methods are typically trained deterministically using human-annotated or synthesized amodal masks, which is not well-suited for amodal segmentation. Though the annotations are carefully annotated by human experts, they represent only one possible solution while other plausible shapes may exist in the invisible region.

For instance, consider the example shown in Fig. 3 (Row 2, 3rd image from left) where two people are standing next to each other and their arms can be at any arbitrary angles. Thus, training a deterministic model using a single amodal ground truth won't capture the stochastic nature of the problem. To overcome this challenge, we propose to formulate the amodal mask completion as a generative process and use a diffusion-based Generative Mask Completion (GMC) module to achieve instance-level mask completion. More specifically, we inherit from the basic structure of the latent-diffusion model proposed in (Rombach et al. 2022). Our model requires the original image, the visible mask of an object, and a condition mask indicating the areas that are possibly under occlusion (which do not need to be very accurate). Using the result from the previous HORI module, we can easily interpolate the condition mask by grouping all the occluders together. With these inputs, the diffusion model is then trained following the standard procedure as in (Rombach et al. 2022).



(a) img (b) gt (c) de-occ (d) ASBU (e) Ours

Figure 4: Visual comparison of the amodal mask completion results from different methods. Though de-occ (c) and ASBU (d) got higher IoU in these cases, Ours (e) generates more sharp and reasonable shapes .

The GMC module largely improves mask completion quality and provides multiple plausible solutions to infer the amodal shapes in the invisible regions. Qualitative comparisons between our approach and existing deterministic approaches are shown in Fig. 4. It is important to note that, though our GMC predictions are visually better, its Intersection-over-Union (IoU) results may not necessarily be higher. For example, in the first row, the output of 'ASBU' has a slightly higher IoU than 'Ours' (82.4 vs 80.1). However, 'Ours' looks more natural, resembling a human stretching an arm, while the 'ASBU' output is not very sensible. This phenomenon applies to the other examples as well: man-made objects, such as the paper bag, the sugar bag, and the chair, should have sharp corners and straight edges, but these cannot be captured by the regular deterministic approaches, resulting in round corners and uneven edges. This urges us to use more advanced metrics to evaluate the plausibility and fidelity of the predicted amodal shapes in the following experiments.

Experiments

Datasets. We conducted extensive experiments to demonstrate the effectiveness of our model on multiple datasets compared with various works. We evaluated our model on three datasets: COCOA, KINS, and ASW. **COCOA** is a subset of COCO2014 and comprises $2,500$ images with $22,163$ instances as the training split and $1,323$ images with $12,753$ instances as the evaluation split. The dataset includes general scenes, such as indoors, outdoors, portraits, and sports events, making it a comprehensive dataset for amodal segmentation. **KINS** is a subset of the KITTI dataset, which is a large-scale traffic dataset. KINS includes $7,474$ images with $95,311$ instances as the training split and $7,517$ images with $92,492$ instances as the testing split. All the scenes in KINS are related to traffic, making it a suitable dataset for testing our model's effectiveness in traffic-related scenarios. **ASW** is our proposed amodal scene in the wild dataset, which includes images collected from daily life situations. The dataset contains $33,049$ images with $316,592$ annotations as the training split and $2,000$ images with 14,969

annotations as the evaluation split. The evaluation split includes $13,240$ occlusion relationships, of which $2,515$ are mutual occlusions. This indicates that mutual occlusion is rather common in daily life situations. More details about ASW dataset are included in the supplementary material.

Training schedule. For our HORI module, we trained on the training split of each dataset and evaluated on their respective evaluation split. We used a batch size of 4 globally, and we trained for $20,000$ iterations on all three datasets. Regarding our GMC diffusion model, we only trained it using the training set of ASW and applied it globally to all three datasets for mask completion. We followed the standard diffusion training schedule, including multiple periods of cosine learning schedule until convergence.

Evaluation metrics. We assessed the accuracy of ordering recovery by calculating the average pairwise accuracy among pairs with occlusion (O-Acc), a commonly-used metric in this area. It should be noted that each pair (e.g., A, B) has two predictions. There are four possible outcomes: A not adjacent to B, A occluding B, B occluding A, and A and B mutually occluding. In addition, we used mean intersection-over-union (mIoU) to evaluate the quality of predicted amodal masks for amodal completion. To ensure the fidelity of the predicted shapes, we measured Fréchet Inception Distance (FID) and Kernel Inception Distance (KID). These metrics are commonly used to evaluate the similarity between the distributions of predicted and ground truth masks. Note that, GMC can predict multiple amodal masks for a given visible instance mask as shown in Fig. 6, for a fair comparison, we fix the prediction to be one.

Main Results

Comparison of the performance on COCOA dataset. Performance comparison of our model and previous methods on the COCOA dataset is presented in Tab. 1. Our HORI module demonstrates the strongest O-Acc performance on COCOA with a score of 90.9%. In terms of mIoU, both HORI-predicted masks and GMC-refined masks exhibit stronger performance than previous methods with a score of 86.89%. However, as noted before, a higher mIoU score may not necessarily indicate better performance. In this case, although both HORI and GMC models have strong mIoU scores, HORI has the lowest fidelity among all methods. This suggests that mIoU may not be the most suitable metric for non-deterministic tasks such as amodal completion. Our GMC model exhibits the strongest fidelity toward the ground truth.

Comparison of the performance on KINS dataset. The KINS dataset solely contains traffic scenes and exhibits a strong inductive bias where larger objects are closer to the camera. Our HORI module can holistically perform ordering recovery and potentially leverage this bias better, whereas previous methods trained pairwise may not capture this information. Therefore, as shown in Tab. 2, HORI outperforms previous methods in ordering recovery by a significant margin. While our GMC module is not specifically trained on traffic-related scenes, it still achieves competitive performance in terms of mask completion quality. However, since the mIoU scores are already high (over 94%) and amodal shapes are relatively uniform in this dataset, the ordering recovery task, rather than mask completion, is the primary factor affecting the overall performance.

Methods	COCOA			
	O-Acc	mIoU	FID	KID
CSDNet	84.7	-	-	-
De-occlusion	87.10	81.35	9.391	0.0034
ASBU	90.33	84.22	-	-
ASBU†	88.00	82.17	8.816	0.0033
HORI (ours)	**90.90**	86.36	12.579	0.0062
GMC (ours)	-	**86.89**	**7.204**	**0.0019**

Table 1: Comparison of ordering recovery and amodal completion on the COCOA dataset. Our method outperforms previous approaches on both tasks. We further evaluated the fidelity of predicted masks to the ground truth masks (smaller is better). † indicates results obtained by retraining using officially released code.

Methods	KINS	
	O-Acc	mIoU
CSDNet	86.4	-
De-occlusion	92.50	94.76
ASBU	92.65	**94.83**
HORI (ours)	**95.22**	93.79
GMC (ours)	-	93.53

Table 2: Comparison of ordering recovery and amodal completion on the KINS dataset. Our O-Acc score outperforms previous methods by a large margin, demonstrating the superior ability of HORI to predict ordering in complex scenes.

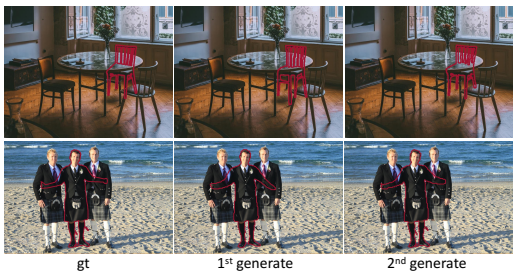
Comparison of the performance on ASW dataset. We re-implement de-occlusion and ASBU on the ASW datasets under the same settings to obtain results for a fair comparison. However, these methods cannot predict mutual occlusion relationships, resulting in a significant performance gap compared to our proposed method (87.27% vs. 82.01% vs. 80.93%) as shown in Tab. 3. Similar to the COCOA result, both HORI-predicted masks and GMC-refined masks exhibit stronger performance in terms of mIoU than previous methods. However, HORI has the poorest fidelity among all methods. By applying GMC, we can obtain mask predictions that are not only good in mIoU but also more sensible predictions. Visualization results on the ASW testing set are shown in Fig. 5. Results in Fig. 6 also show GMC is capable of generating multiple reasonable predictions.

Discussions

Ablations on the structure of HORI. The proposed HORI module includes multiple novel structural designs, which are ablated one-by-one in Tab. 4. By simply using masked attention to encode visible instance masks to the model structure, the O-Acc score reaches 85.6. The 'ins concat' operation involves concatenating visible instance masks with feature maps before amodal prediction. 'Matrix inference' uses the predicted occlusion relationship matrix instead of amodal predicted masks to interpolate the final occlusion relation-

(a) image (b) gt (c) De-occ (d) ASBU (e) Ours

Figure 5: Amodal mask completion results on ASW evaluation set. Compared with existing methods, our approach generates more visually plausible amodal shapes. The inferred amodal masks are more consistent with the ground truth. Particularly for man-made objects, our GMC module generates sharper edges and corners with higher quality.



gt 1st generate 2nd generate

Figure 6: Amodal completion results showcasing GMC's capability of generating multiple reasonable predictions.

Methods	ASW			
	O-Acc	mIoU	FID	KID
De-occlusion	80.93	88.32	4.466	0.0012
ASBU	82.01	88.84	4.438	0.0011
HORI (ours)	**87.27**	90.24	4.946	0.0022
GMC (ours)	-	**90.53**	**4.046**	**0.0009**

Table 3: Comparison of ordering recovery and amodal completion on the ASW dataset. Beyond the mIoU score, our model also achieves better FID and KID, indicating the superior ability to generate high-quality amodal masks.

ship. 'Upsample inputs' involves scaling the image's shortest edge to 1024 during inference. These operations all have a positive impact on O-Acc.

Efficiency on object occlusion reasoning. As we proposed a novel occlusion relation matrix to infer the occlusion relationship among all the instances, our method is more efficient than the pair-wise relation reasoning framework (Nguyen and Todorovic 2021; Zhan et al. 2020b), especially for complicated real-world scenarios. Our end-to-end method only requires one forward pass to infer all the occlusion correlations among the instances in the images. In contrast, previous methods require a complicated pipeline by first segmenting all the instances, then inputting every instance pair into the relation reasoning network. Hence,

MA	IC	MI	UI	O-Acc
✓				85.6
✓	✓			87.2
✓	✓	✓		89.4
✓	✓	✓	✓	90.9

Table 4: Ablations for the HORI module on COCOA dataset. O-Acc is reported here. As different components are added to the HORI module, the performance gradually improves. MA refers to Masked Attention, IC refers to Ins Cancat, MI refers to Matrix Inference and UI refers to Upsample Inputs.

the computational cost and the inference time will increase linearly according to the number of occlusion pairs. Taking COCOA as an example, there are 1323 images with 22630 occlusion pairs. We test the inference speed on a single 3090Ti GPU card for our framework and the pair-wise reasoning framework (Nguyen and Todorovic 2021). We achieve an average inference speed of 5 FPS to get the occlusion relationship while ASBU (Nguyen and Todorovic 2021) only achieves 1.68 FPS.

Conclusion

In this paper, we propose a novel framework for amodal scene analysis that comprises the Holistic Occlusion Relation Inference (HORI) module and the Generative Mask Completion (GMC) module. The HORI module predicts an occlusion relationship matrix in a single pass, which largely improves the inference efficiency and enables reasoning for mutual occlusion. The GMC module formulates amodal mask completion as a generative process and provides multiple high-quality plausible solutions. Our experimental results on COCOA, KINS, and the proposed ASW benchmark demonstrate state-of-the-art performance and robustness to various occlusion scenarios. Our framework and benchmark can serve as essential baselines for future amodal scene analysis research, with potential applications in robotics, autonomous driving, and image editing.

Acknowledgments

Y. Liu acknowledges the support of start-up funding from The University of Adelaide for their participation in this work. We express our gratitude to The University of Adelaide High-Performance Computing Services for providing the GPU Compute Resources, and to Mr. Wang Hui and Dr. Fabien Voisin for their valuable technical support.

This work stands as a testament to the dedication and expertise of https://yifaninmemory.vmv.re/ Dr. Yifan Liu. We are deeply grateful to Dr. Liu for her unparalleled contribution and leadership throughout this project. Her insights and guidance have been invaluable. We wish to honour and remember Dr. Yifan Liu for her enduring impact and legacy in this field.

References

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 1290–1299.

Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2022. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*.

Dhariwal, P.; and Nichol, A. Q. 2021. Diffusion Models Beat GANs on Image Synthesis. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *NeurIPS*.

Ehsani, K.; Mottaghi, R.; and Farhadi, A. 2018. Segan: Segmenting and generating the invisible. In *CVPR*, 6144–6153.

Follmann, P.; König, R.; Härtinger, P.; Klostermann, M.; and Böttger, T. 2019. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1328–1336. IEEE.

Gilroy, S.; Jones, E.; and Glavin, M. 2019. Overcoming occlusion in the automotive environment—A review. *IEEE Transactions on Intelligent Transportation Systems*, 22(1): 23–35.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *NIPS*.

Ho, J.; Saharia, C.; Chan, W.; Fleet, D. J.; Norouzi, M.; and Salimans, T. 2022. Cascaded Diffusion Models for High Fidelity Image Generation. *J. Mach. Learn. Res.*, 23(47): 1–33.

Hoiem, D.; Stein, A. N.; Efros, A. A.; and Hebert, M. 2007. Recovering occlusion boundaries from a single image. In *ICCV*, 1–8. IEEE.

Hsiao, E.; and Hebert, M. 2014. Occlusion reasoning for object detectionunder arbitrary viewpoint. *IEEE transactions on pattern analysis and machine intelligence*, 36(9): 1803–1815.

Jain, J.; Li, J.; Chiu, M.; Hassani, A.; Orlov, N.; and Shi, H. 2022. OneFormer: One Transformer to Rule Universal Image Segmentation. *arXiv preprint arXiv:2211.06220*.

Jiang, Z.; Liu, B.; Schulter, S.; Wang, Z.; and Chandraker, M. 2020. Peek-a-boo: Occlusion reasoning in indoor scenes with plane representations. In *CVPR*, 113–121.

Jolicoeur-Martineau, A.; Piché-Taillefer, R.; des Combes, R. T.; and Mitliagkas, I. 2020. Adversarial score matching and improved sampling for image generation.

Kang, S. B.; Szeliski, R.; and Chai, J. 2001. Handling occlusions in dense multi-view stereo. In *CVPR*, volume 1, I–I. IEEE.

Kar, A.; Tulsiani, S.; Carreira, J.; and Malik, J. 2015. Amodal completion and size constancy in natural scenes. In *ICCV*, 127–135.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2022. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*.

Kimia, B. B.; Frankel, I.; and Popescu, A.-M. 2003. Euler spiral for shape completion. *International journal of computer vision*, 54(1-3): 159–182.

Kortylewski, A.; Liu, Q.; Wang, H.; Zhang, Z.; and Yuille, A. 2020. Combining compositional models and deep networks for robust object classification under occlusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1333–1341.

Li, F.; Zhang, H.; Liu, S.; Zhang, L.; Ni, L. M.; Shum, H.-Y.; et al. 2022a. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint arXiv:2206.02777*.

Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2022b. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479: 47–59.

Li, K.; and Malik, J. 2016. Amodal instance segmentation. In *European Conference on Computer Vision*, 677–693. Springer.

Ling, H.; Acuna, D.; Kreis, K.; Kim, S. W.; and Fidler, S. 2020. Variational amodal object completion. *NeurIPS*, 33: 16246–16257.

Lu, R.; Xue, F.; Zhou, M.; Ming, A.; and Zhou, Y. 2019. Occlusion-shared and feature-separated network for occlusion relationship reasoning. In *ICCV*, 10343–10352.

Mohan, R.; and Valada, A. 2022. Amodal panoptic segmentation. In *CVPR*, 21023–21032.

Nanay, B. 2018. The importance of amodal completion in everyday perception. *i-Perception*, 9(4): 2041669518788887.

Nguyen, K.; and Todorovic, S. 2021. A weakly supervised amodal segmenter with boundary uncertainty estimation. In *ICCV*, 7396–7405.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Qi, L.; Jiang, L.; Liu, S.; Shen, X.; and Jia, J. 2019. Amodal instance segmentation with kins dataset. In *CVPR*, 3014–3023.

Ren, X.; Fowlkes, C. C.; and Malik, J. 2006. Figure/ground assignment in natural images. In *ECCV*, 614–627. Springer.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022a. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022b. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Silberman, N.; Shapira, L.; Gal, R.; and Kohli, P. 2014. A contour completion model for augmenting surface reconstructions. In *ECCV*, 488–503. Springer.

Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics.

Sun, Y.; Kortylewski, A.; and Yuille, A. 2022. Amodal Segmentation Through Out-of-Task and Out-of-Distribution Generalization With a Bayesian Model. In *CVPR*, 1215–1224.

Tighe, J.; Niethammer, M.; and Lazebnik, S. 2014. Scene parsing with object instances and occlusion ordering. In *CVPR*, 3748–3755.

Tran, M.; Vo, K.; Yamazaki, K.; Fernandes, A.; Kidd, M.; and Le, N. 2022. AISFormer: Amodal Instance Segmentation with Transformer. *arXiv preprint arXiv:2210.06323*.

Vahdat, A.; Kreis, K.; and Kautz, J. 2021. Score-based Generative Modeling in Latent Space. In *Advances in neural information processing systems*.

Wang, P.; and Yuille, A. 2016. Doc: Deep occlusion estimation from a single image. In *ECCV*, 545–561. Springer.

Wu, J.; Tenenbaum, J. B.; and Kohli, P. 2017. Neural scene de-rendering. In *CVPR*, 699–707.

Xiao, Y.; Xu, Y.; Zhong, Z.; Luo, W.; Li, J.; and Gao, S. 2021. Amodal segmentation based on visible region segmentation and shape prior. In *AAAI*, volume 35, 2995–3003.

Yamaguchi, K.; McAllester, D.; and Urtasun, R. 2014. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *ECCV*, 756–771. Springer.

Yan, X.; Wang, F.; Liu, W.; Yu, Y.; He, S.; and Pan, J. 2019a. Visualizing the invisible: Occluded vehicle segmentation and recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7618–7627.

Yan, X.; Wang, F.; Liu, W.; Yu, Y.; He, S.; and Pan, J. 2019b. Visualizing the invisible: Occluded vehicle segmentation and recovery. In *ICCV*, 7618–7627.

Zeng, Y.; Lin, Z.; Zhang, J.; Liu, Q.; Collomosse, J.; Kuen, J.; and Patel, V. M. 2022. SceneComposer: Any-Level Semantic Image Synthesis. *arXiv preprint arXiv:2211.11742*.

Zhan, X.; Pan, X.; Dai, B.; Liu, Z.; Lin, D.; and Loy, C. C. 2020a. Self-supervised scene de-occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3784–3792.

Zhan, X.; Pan, X.; Dai, B.; Liu, Z.; Lin, D.; and Loy, C. C. 2020b. Self-supervised scene de-occlusion. In *CVPR*, 3784–3792.

Zheng, C.; Dao, D.-S.; Song, G.; Cham, T.-J.; and Cai, J. 2021. Visiting the Invisible: Layer-by-Layer Completed Scene Decomposition. *IJCV*, 129(12): 3195–3215.

Zhu, H.; Tang, P.; Park, J.; Park, S.; and Yuille, A. 2019. Robustness of object recognition under extreme occlusion in humans and computational models. *arXiv preprint arXiv:1905.04598*.

Zhu, Y.; Tian, Y.; Metaxas, D.; and Dollár, P. 2017. Semantic amodal segmentation. In *CVPR*, 1464–1472.