# CondDiff-AMO: Integrating Conditional Diffusion Mechanism for Unified Amodal Mask Generation

**CaiJie Zhao**[1], **Bob Zhang**[1,2*]

[1] Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Taipa, Macau
[2] Centre for Artificial Intelligence and Robotics, Institute of Collaborative Innovation, University of Macau, Taipa, Macau
yc47911@um.edu.mo, bobzhang@um.edu.mo

## Abstract

Aiming to estimate the full extent of partially occluded objects, amodal segmentation is a critical capability for visual intelligence. Existing methods suffer from limitations in efficiency and precision, due to their reliance on auxiliary information or two-stage architectures. Furthermore, they lack generalizability, failing to meet practical requirements. To overcome these challenges, we proposed a new paradigm, CondDiff-AMO, that interprets amodal segmentation as a denoising problem by leveraging diffusion models. Methodologically, the designed novel framework consists of three key innovations to adapt the task characteristics and unlocks the diffusion models' potential in amodal segmentation, including a masking strategy in the forward process, an adaptive transformer for conditional feature extraction, and visual-guided sampling. In the forward process, progressive masking strategy converts ground-truth masks to visible masks, simulating amodal segmentation process to enhance reasoning regarding occluded areas. For architectural design, a pyramid network with feature refinement extracts adaptive and representative conditional priors, improving the guidance in the denoising process of diffusion models. As for the sampling stage, a visible mask is incorporated with an ensemble strategy, restricting the prediction on occluded part. Experiments were conducted on five well-known datasets under supervised and zero-shot learning, with the results confirming that CondDiff-AMO outperforms state-of-the-art methods.

**Code** —
https://github.com/Carinazhao22/CondDiff-AMO.git

## 1 Introduction

Due to the prevalence of the occlusion phenomenon in real-world scenarios, varying from the cluttered natural environment to urban settings, occlusion comprehension is one of the significant capabilities of visual intelligence (Palmer 1999). To achieve this goal, amodal segmentation was introduced to infer the full shape of an object, involving both visible and invisible portions (Li and Malik 2016). Various powerful models, such as Mask-RCNN (He et al. 2017), Mask2Former (Cheng et al. 2022), and Seg-Anything (SAM) (Kirillov et al. 2023), have achieved remarkable progress in modal segmentation, i.e., only segments visible pixels of an instance, while they fail to imitate the human ability to predict the full extent of partially occluded objects, leading to fragment object segmentation. Amodal segmentation has been explored in various downstream tasks to further boost their performance, including 3D reconstruction from a single image (Kanazawa et al. 2018; Kar et al. 2015; Ozguroglu et al. 2024; Wu et al. 2023; Zou and Hoiem 2020) aided by complete geometric priors, autonomous driving (Breitenstein and Fingscheidt 2022; Qi et al. 2019; Wada, Okada, and Inaba 2019) with traffic risks reduced, and robotic gripping systems (Inagaki et al. 2019; Wada et al. 2018) with stable gripping.

Despite the crucial advances, these previous methods suffer from suboptimal results generated by fixed-parameter networks (Yao et al. 2022; Li et al. 2022; Xiao et al. 2021; Gao et al. 2023), error accumulation from two-stage architectures (Xiao et al. 2021; Gao et al. 2023) and all are restricted by supervised settings. To address these issues, Pix2gestalt (Ozguroglu et al. 2024), is designed for amodal completion with amodal mask as secondary outputs but without specifically designed for segmentation. In light of this situation, a natural question arises: *How do we integrate the characteristics of amodal segmentation into diffusion models for building a zero-shot amodal segmentation model?*

Recently, diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2020; Song and Ermon 2019) have been applied in general modal segmentation, among which conditional diffusion models are prime to be mainly adopted. Conditional diffusion models possess powerful generative capabilities, as they employ a step-by-step denoising method with conditioning to recover suboptimal masks in each step. Under this mechanism, conditional diffusion models achieve the optimal solution ultimately, which is vital for amodal segmentation. Accordingly, we expect to exploit the impressive generative capabilities of diffusion models to facilitate amodal segmentation in open-world settings for meeting the practical requirements.

To this end, unique characteristics of amodal segmentation should be adapted into diffusion models targeted at the forward process, training and sampling strategies. Amodal segmentation predicts amodal masks based on clues from modal masks, while traditional Gaussian noises provide

*Corresponding Author

pixel-level corruption and fail to simulate the inferring process from modal masks to amodal masks. Amodal segmentation requires occlusion and context information from modal masks and images to infer amodal masks, but conventional diffusion model did not extract conditional features from binary masks and images, together with time embedding. Amodal segmentation demands reasonable inference based on information from modal masks instead of random sampling with the traditional strategy. We notice that directly applying a pre-trained diffusion model and strategy in amodal segmentation without considering its uniqueness.

In response to the above problems, we designed an innovative interpretation of amodal segmentation by representing it as a denoising process. It can be implemented through a series of denoising diffusion steps with visible segmentation masks being the noisy version of the ground truth. Following this paradigm, we proposed a new framework called CondDiff-AMO, sufficiently leveraging the generative ability of a diffusion model and adapting it to the amodal segmentation scenario. Firstly, to improve the masking process, we introduced a novel transition-state masking strategy in the forward process called Unidirectional State Transition (UST) to generate noisy ground truth masks, where each pixel operates a unidirectional random state transition to visible masks. This strategy progressively transfers ground truth into its noisy version (modal masks), imitating the inferring process of amodal segmentation for later denoising process. Secondly, we designed Pyramid Condition Network (PCN) to effectively encode hierarchical condition features from images, modal masks and time embedding. Denoising Network (DN) is proposed to fuse multi-level conditioned features from PCN for the later prediction. Thirdly, we introduced Visible-Guided Prediction Ensemble (VGPE) to combine predictions from different sampling steps guided by modal masks that is vital for generating reasonable masks.

To evaluate the superiority of the proposed approach, we conducted two kinds of experiments, closed-world and zero-shot experiments, on five prominent datasets. The results show that compared to zero-shot methods along with the supervised approaches, CondDiff-AMO attains a significant enhancement in amodal segmentation. Our contributions can be summarized as:

- We proposed a diffusion-based method for amodal segmentation called CondDiff-AMO, which is the first model (to the best of our knowledge) using an amodal mask as the starting point to treat the amodal segmentation task as a denoising paradigm.
- In terms of the task-specific characteristics of amodal segmentation, a suitable architecture incorporating a unidirectional state transition masking strategy, a Pyramid Condition Network with a denoising network, and the visible-restricted prediction ensemble, was elaborately designed to unleash the potential amodal capabilities of the diffusion model.
- Our proposed method CondDiff-AMO surpassed existing supervised and zero-shot approaches in supervised and open-world settings, reaching state-of-the-art performance on five benchmarks.

# 2 Related Work

## 2.1 Amodal Segmentation

Amodal segmentation aims to estimate the full shape of the partially occluded objects. The prior work can be classified into two groups: 1) shape-prior models (Yao et al. 2022; Xiao et al. 2021; Gao et al. 2023; Zhang et al. 2024) and 2) diffusion-based models (Ozguroglu et al. 2024). Techniques in the first group incorporated extra information to boost performance but reduced the convenience. Another line of work focuses on two-stage training, suffering from error-accumulation, where occlusion relation or shape prior were constructed to aid later completion. Another limitation held by the first-group methods is that fixed-parameter networks fail to consider diverse categories and complicated occlusions, resulting in suboptimal completion.

The second kind of approaches, driven by diffusion models, address the above issues. SDAmodal (Zhan et al. 2024) employed a pre-trained Stable Diffusion model to recover amodal masks. Pix2gestalt (Ozguroglu et al. 2024) utilized a latent diffusion framework to perform zero-shot amodal segmentation. Those methods take advantage of diffusion models to show superior performance under zero-shot learning over the approaches in the first group, while they did not consider the characteristics of amodal segmentation for fitting into diffusion models, which degrade the performance.

## 2.2 Diffusion Models

Diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2020; Song and Ermon 2019) aim to gradually denoise data samples from random noise via parameterized Markov chains, initially applying this innovative approach to image generation with notable success. Due to its powerful generative ability, it has been extended to other fields such as deblurring (Whang et al. 2022; Lee et al. 2022), camouflaged object detection (Chen, Sun, and Lin 2024), and image segmentation (Wu et al. 2024; Baranchuk et al. 2021; Amit et al. 2021), accomplishing state-of-the-art performance in their benchmarks. Building on successful experiences in previous literature, this paper explores the possibility of diffusion models in amodal segmentation to solve above limitation.

# 3 Methods

In this section, we illustrate our CondDiff-AMO framework, which gradually masks the amodal mask to its noisy version (modal mask) in the forward process and then makes predictions by conditioning each step with images and masks prior in the backward process. As presented in Fig. 1, CondDiff-AMO framework contains a Pyramid Condition Network (PCN) and Denoising Network (DN). The PCN aims to extract features from images, masks and together with time tokens, while DN is employed to eliminate noises conditioning on fused features and formulate predictions. During the training stage, our model is optimized to denoise noisy masks that are corrupted by Unidirectional State Transitions (UST) specially designed in forward process for simulating amodal segmentation tasks. In the sampling stage, the proposed model progressively denoises sampled noise led by modal masks, forming the optimal prediction.

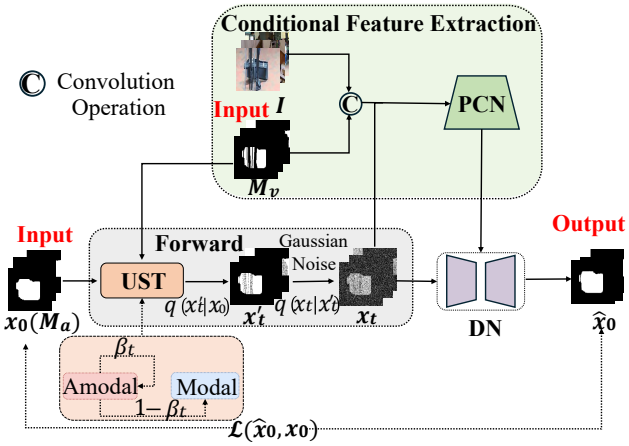


Figure 1: Illustration of our CondDiff-AMO framework. An amodal mask $x_0$ ($M_a$) and a modal mask $M_v$ are subject to Unidirectional State Transition (UST) and Gaussian noise to construct the noisy mask $x_t$ at time $t$. Pyramid Condition Network (PCN) generates hierarchical and adaptive features from image $I$, $M_v$, $x_t$ and $t$ and subsequently, the prediction $\hat{x}_0$ generated from Denoising Network (DN) is optimized by minimizing the loss between the ground truth and prediction, represented as $\mathcal{L}(\hat{x}_0, x_0)$.

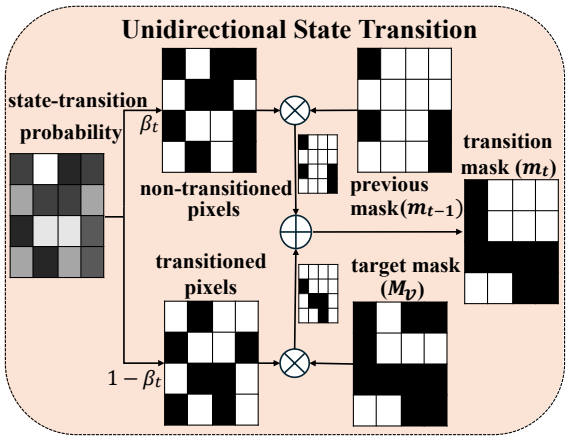


Figure 2: The illustration of a transitional sample $m_t$ from Unidirectional State Transition (UST) model in our CondDiff-AMO.

### 3.1 Background and Task Definition

The proposed CondDiff-AMO relies on diffusion models, including forward and reverse processes. The forward process $q(x_{1:T} \mid x_0)$ uses a Markov chain to gradually transform the data distribution $x_0 \sim q(x_0)$ into its complete noisy version $\{x_t\}_{t=1}^T$, while the reverse process deploys a gradual denoising procedure $p_\theta(x_{0:T})$ to transform a random noise back into the original data distribution. The forward process to construct $\{x_t\}_{t=1}^T$ is presented as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where $t$ runs from 1 to $T$, and mean and variance of the forward process are controlled by noise schedule $\beta_t \in (0,1)$. The reverse process utilizes a neural network $f_\theta$ to perform a sequence of denoising operations, starting from $p(x_T) = \mathcal{N}(x_T; 0, \mathbf{I})$, to transform back the clean mask. The reverse distribution obtained by the network is formulated as:

$$p(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t)). \tag{2}$$

Amodal segmentation segments the complete shapes of the partially occluded object, while general modal segmentation only contains the visible region. Zero-shot amodal segmentation represents adapting trained models in new datasets without fine-tuning and supervised amodal segmentation utilizes the same datasets for training and testing. For each instance, the model takes an image $I$ and a visible mask (model mask) $M_v$ as the clues to output a comprehensive amodal mask (full mask) $M_a$, involving both visible and occluded regions. The objective is to estimate the $M_a$ leveraging the features from both $I$ and $M_v$. To accomplish this, we utilize conditional diffusion models to condition images and binary mask features for predicting $M_a$. In our proposed CondDiff-AMO, we set ground truth $M_a$ as $x_0$ and final denoised mask $\hat{x}_0$ as the prediction $\hat{M}_a$. A network $f_\theta(x_t, I, M_v, t)$ is trained to predict the denoised mask based on the clues from image $I$ and modal mask $M_v$. Category labels are not required. Followed (Chen, Sun, and Lin 2024), $\sum_\theta(x_t,t)$ is set to $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$ and $\mu_\theta(x_t,t)$ is:

$$\mu_\theta(x_t,t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{x}_0, \tag{3}$$

where $\hat{x}_0$ is estimated by our model $f_\theta(x_t, I, M_v, t)$.

### 3.2 Unidirectional State Transition

Existing diffusion models apply Gaussian noises to generate pixel-level corrupted data directly from the ground truth, which is not suitable for the amodal segmentation, restoring missing part to infer the amodal masks based on the visible parts. Inspired by (Wang et al. 2023) to explore the diffusion model in our task, we propose Unidirectional State Transition (UST), to convert the pixel states for creating the noisy masks $\{x_t\}_{t=1}^T$. During the forward process, we progressively degrade $M_a$, transiting it into the visible mask $M_v$. This means that we have $x_0 = M_a$ and $x_T = G(M_v)$ with Gaussian noise represented as $G$, and the intermediate mask $m_t$ is a transitional phase between $M_a$ and $M_v$ at any intermediate timestep $t \in \{1, 2, \ldots, T-1\}$. The UST corrupts the noisy mask more severely, containing more missing parts, as time increases. As demonstrated in Fig. 2, the UST utilizes the previous mask $m_{t-1}$, $M_v$ and a state-transition probability as input and then generates a transitional mask $m_t$. Specifically, the UST employs Gumbel-max sampling to gain the transitioned pixels based on the state-transition probability and then the transferred pixels will take values from $M_v$ and the non-transferred pixels stay unchanged. The $m_t$ is also added with Gaussian noise.

Mathematically, we utilize, $x_t^{i,j}$, a one-hot vector to specify the state of pixel $(i,j)$ in $m_t$ and set $x_0^{i,j} = [1,0]$ and $x_T^{i,j} = [0,1]$ to illustrate the amodal state and modal state,

respectively. The UST can be formulated as:

$$q(x_t^{i,j}|x_{t-1}^{i,j}) = x_{t-1}^{i,j} P_t, \text{ with } P_t = \begin{bmatrix} \beta_t & 1-\beta_t \\ 0 & 1 \end{bmatrix} \quad (4)$$

where $\beta_t \in [0,1]$, and $1-\beta_t$ represents the transition probability used in UST. $P_t$ restricts all pixels in the amodal state never convert back to the modal state for $q(x_t|[0,1]) = [0,1]$. Eq. 3 represents the forward process $q(x_t|x_{t-1})$ to gradually transfer the data distribution $x_0 \sim q(x_0)$ into noise $x_t$. Based on Eq. 4, the marginal distribution of $x_t$ is formulated as:

$$q_t(x^{i,j}|x_0^{i,j}) = x_0^{i,j} \prod_{i=1}^{t} P_i = x_0^{\bar{P}_t} = x_0 \begin{bmatrix} \bar{\beta}_t & 1-\bar{\beta}_t \\ 0 & 1 \end{bmatrix}, \quad (5)$$

where $\bar{\beta}_t = \prod_{i=1}^{t} \beta_i$. The forward process incorporated with the UST empowers the CondDiff-AMO to consider the task characteristics and leverage the capabilities of diffusion models.

### 3.3 Pyramid Condition Network and Denoising Network

In contrast to other traditional models either accessing extra information or applying two-stage training, CondDiff-AMO employs conditional diffusion model to obtain prior knowledge in $I$ and $M_v$. In detail, we design a transformer-based network called Pyramid Condition Network (PCN) presented in Fig. 3A, to extract hierarchical features from $I$ and $M_v$ as conditions, which are then assisting the downstream denoising process in the Denoising Network (DN) shown in Fig. 3B. We state two objectives in the design of our architecture: 1) the PCN extracting more discriminative and multi-level features from image and binary mask, embedding with time tokens, 2) the DN effectively integrating multi-level features for aiding the denoising step by adaptively supplying conditional features. We introduce the details of the PCN and DN in the following sections.

**Pyramid Condition Network.** PCN aims to obtain representative features from images and masks, containing background information of the image as well as the characteristics of the object itself, which plays a crucial role to reason about the invisible portion in amodal masks. Based on the uniqueness of amodal segmentation, the design of the PCN should address three primary challenges: 1) enabling feature extraction on images and modal masks without destroying the original network structure and pre-trained parameters, 2) forcing the model concentrate on global information when the noise mask is extremely corrupted as time increases, and shift attention to detailed areas when the missing part has been recovered gradually in the mask.

To address these challenges, the PCN utilizes Pyramid Vision Transformer (PVT) (Wang et al. 2022b) as the backbone to extract hierarchical features that are fused together in the DN. Image $I$ and visible mask $M_v$ are concatenated together to form a four-channels input and a simple convolution layer is then employed to adjust the four-channels to three-channels inputs defined as $X_{input}$, preparing for the backbone extraction. As depicted in Fig. 3, the PVT extracts multi-level features stated as $\{F_i\}_{i=1}^4$ from $X_{input}$ with $x_t$ and $t$ embedding. The PVT layer, consisting of a Patch Embedding module and a Transformer Encoder in conjunction

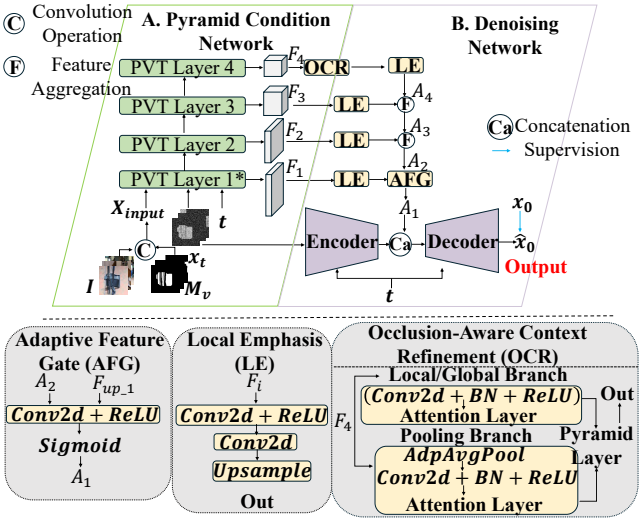


Figure 3: The details of the subnetworks in our CondDiff-AMO. It consists of two networks: A. Pyramid Condition Network (PCN) and B. Denoising Network (DN). The PCN extracts hierarchical features as conditions to guide the DN in restoring clean mask prediction from noisy mask. Zoom in for better visualization.

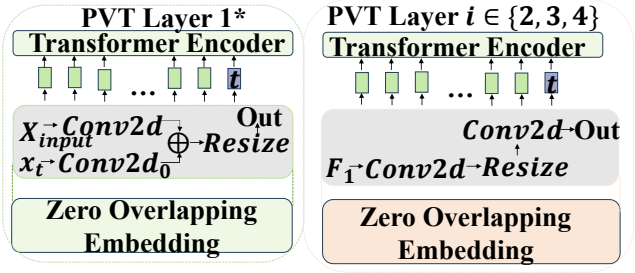


Figure 4: The layer details of the Pyramid Condition Network. Zoom in for better visualization.

with a multi-level mechanism, ensures a comprehensive fusion of the conditioned features from coarse-grained to fine-grained at each stage. Amodal segmentation demands identifying occluded object regions by combining local fined details (e.g., boundaries) and global semantic context (e.g. object shape prior), thus requiring feature fusion from $\{F_i\}_{i=1}^4$. Mathematically, the procedure is formed as:

$$X_{input} = Conv2d(Cat(I, M_v)) \quad (6)$$

$$\{F_i\}_{i=1}^4 = PVT_i(x_t, X_{input}, t), \quad (7)$$

where $Conv2d$ and $Cat$ is defined as a convolution layer and a concatenation operation. $x_t$ represents the noisy mask at time $t$.

However, the deepest features, $F_4$, from PVT captures high-level semantic information but suffers from limited local details and reduced spatial resolution, causing significant performance degradation in predicting occluded parts and misalignment with shallow features. Inspired by Atrous Spatial Pyramid Pooling (ASPP) (Chen et al. 2017), we design

a module called **Occlusion-Aware Context Refinement** to enrich spatial details in $F_4$ for matching the resolution of shallow features and highlight occluded-relevant areas to prioritize occluded regions during the feature fusion. Furthermore, applying images and visible masks as input and using the same features in the reverse process fails to consider various stages requiring different requisite information for the denoising. The PCN should emphasize the overall image when the Signal-to-Noise Ratio of the noise mask is low, while switch focus to the local regions once the missing parts of the mask have been recovered (Chen, Sun, and Lin 2024). We adopt Zero Overlapping Embedding and Time Token Concatenation (Sun et al. 2025), as shown in Fig. 4, to adaptively immerse the noise mask $x_t$ and time $t$, resulting in corresponding requisite features applied in varying stages.

**Occlusion-Aware Context Refinement (OCR).** ASPP contains parallel atrous convolutions with fixed dilation rates (e.g., 6,12,18) and a global branch. They are fused via a simple convolution to reduce dimension without designing mechanisms to prioritize specific areas. Thus, ASPP is not appropriate in complex tasks like amodal segmentation (inferring hidden objects). To improve the deepest features for features fusion and task-specific traits, we propose OCR shown in the lower part of Fig. 3, which defines three branches cooperating with pyramid attention in this submodule. In detail, the Local Branch and Global Branch employs small dilation rates (1 and 3) and larger dilation rates (6 and 12) to gain fine-grained textures (e.g., occlusion boundaries) and long-range dependencies, respectively. Pooling Branch provides high-level context by aggregating global scene semantics. Each branch includes a channel attention layer to reweight channels for emphasizing occlusion-relevant features. The fusion of the output from three branches to enrich deepest features in the PVT and then pyramid attention performs spatial attention fusion with multi-scale sampling. The pyramid attention simulates the occlusion at varying sizes and dynamically reweights regions critical for occlusion reasoning.

**Denoising Network (DN).** In the reverse process, we employ a simple U-shape structure as DN to estimate the denoised mask at each $t$ step with input containing of time $t$, pyramid feature maps $\{F_i\}_{i=1}^4$ obtained from PCN and OCR submodule, and the noisy mask $x_t$. Instead of employing complex decoders, we utilize simple structures to achieve competitive performance by leveraging the benefits of iterative denoising processes in diffusion models. Specifically, the Local Emphasis (LE) (Wang et al. 2022a) is adopted to regulate the multi-scale features $\{F_i\}_{i=1}^4$ to the same size $\frac{H}{4} \times \frac{W}{4}$, represented as $\{F_{up\_i}\}_{i=1}^4 = \text{LE}\left(\{F_i\}_{i=1}^4\right)$, then passed through a convolution for gradually fusing them into $\{A_i\}_{i=1}^4$. Considering the task traits, we design a feature gate, **Adaptive Feature Gate** (AFG), to integrate the previous fused feature $A_2$ and $F_{up\_1}$ for forming the final conditioned feature $A_1$ used in the decoder. Conventional linear fusion uses fixed convolution operations to merge features, assuming features high-level context $A_2$ and low-level details $F_{up\_1}$ contributing equally to the final feature map. The linear fusion merges them without distinguishing between the background noise and useful features. Amodal segmentation requires the network prioritize $A_2$ in global information for locating occlusion regions and $F_{up\_1}$ in boundary regions to refine hidden details. We introduce AFG to address this issue by suppressing irrelevant background textures and amplifying the structural features about hidden part. The AFG is represented as:

$$Gate = Sigmoid(Conv2d(CR(A_2, F_{up\_1}))), \tag{8}$$

where $Conv2d$ is defined as convolution operation and $CR$ represents the combination of convolution and ReLU. Thus, the procedure of fusion is described as:

$$\begin{aligned} A_4 &= F_{up\_4}, \\ A_i &= Conv2d(A_{i+1}, F_{up\_i}), \quad i \in \{2, 3\}, \\ A_1 &= Gate * A_2 + (1 - Gate) * F_{up\_1}, \end{aligned} \tag{9}$$

where $Gate = AFG(A_2, F_{up\_1})$ and $Conv2d$ represent convolution operators to linearly fuse features. Finally, we adopt a simple encoder and decoder to obtain discriminative features from the input at t and then make the prediction, $\hat{x}_0$, conditioning in $A_1$. We optimize $\mathcal{L}(\hat{x}_0, x_0)$ to train our model, which is formulated as:

$$\mathcal{L}(\hat{x}_0, x_0) = \mathcal{L}_{BCE}^w(\hat{x}_0, x_0) + \mathcal{L}_{IoU}^w(\hat{x}_0, x_0). \tag{10}$$

$\mathcal{L}_{BCE}^w$ and $\mathcal{L}_{IoU}^w$ indicate weighted binary cross entropy (BCE) and weighted intersection-over-union (IoU) loss.

### 3.4 Visible-Guided Prediction Ensemble

Our denoising procedure, like conventional denoising model, performs incremental denoising steps for sampling $x_T$ from a standard normal distribution over $T$ iterative steps. During the stepwise denoising process, the divergence between the ground truth and the predicted mask is progressively narrowed, ultimately yielding a refined and accurate outcome. We observe that the predicted mask generated over each step would provide valuable insight for final prediction and the sampling process should leverage the visible portion to build the more appropriate and reasonable amodal mask. Thus, we design Visible-Guided Prediction Ensemble to sample the denoised mask $\hat{x}_0$, which integrates denoised masks from each step to refine the reliability and precision of the final prediction and applies modal masks $M_v$ to restrict the visible part in the later half-part of $T$.

We design a mechanism called **Visible-Guided Prediction Ensemble** (VGPE) to combine the denoised masks over sampling steps restricted by modal masks for handling uncertainty in ambiguous regions. VGPE aims to explicitly guide the model preserving visible regions while allowing flexibility in occluded areas by directing the model's generative capacity to focus on occluded regions. Specifically, after generating the prediction $\hat{x}_{0\_raw}$ from step $t$, we replace the visible portion in $\hat{x}_{0\_raw}$ with $M_v$ and this adjustment is applied during the latter half of the sampling steps to ensure the predictions are sufficiently stable to incorporate hard constraints in non-occluded parts and diffuse in occluded regions. Formally,

$$\hat{x}_0 = \begin{cases} \hat{x}_{0\_raw}, & t < \frac{T}{2}, \\ \hat{x}_{0\_raw} * (1 - M_v) + M_v * 1.0, & t \geq \frac{T}{2}. \end{cases} \tag{11}$$

Each intermediate prediction $\hat{x}_0$ sampling from $T$ obtains insightful information for the final predictions, and thus we

apply the strategy from (Chen, Sun, and Lin 2024) to further improve the predicted mask. For each step $t$, the denoised image $\hat{x}_0$ is recorded to form a sequence of prediction $\{P_t\}_{t=1}^{T}$. The pixel's final probability is determined by averaging the values across binary mask $\{P_b^t\}_{t=1}^{T}$ via the thresholding, and the probability of each point is the mean of all predictions over $T$, shown as below:

$$P_{final} = \left| \frac{\sum_{t=1}^{T} P_b^t}{T} + \frac{1}{2} \right| * mean(P_t). \tag{12}$$

# 4 Experiments

## 4.1 Training Details

**Datasets**. To evaluate the superiority of our proposed CondDiff-AMO, we conducted comprehensive experiments either restricted by supervised settings (training and testing on same datasets) or zero-shot experiments on five amodal segmentation benchmarks. To achieve sufficient training in zero-shot experiments, a large-scale and high-quality dataset is necessary. We utilized the Pix2gestalt dataset for training, following the same procedure as other advanced methods. The model trained in Pix2gestalt dataset is then examined on five benchmarks, KINS (Qi et al. 2019), COCOA (Zhu et al. 2017), COCOA-cls(Ehsani, Mottaghi, and Farhadi 2018), D2SA (Follmann et al. 2019), and MP3D (Zhan et al. 2024). They are also applied in the supervised experiments.

**Evaluation Metrics**. We employed mean intersection-over-union (mean-IoU) as a metric to evaluate the quality of predicted amodal masks. Following community convention, we measured mIoU$_{full}$ between the predicted masks and the ground-truth amodal masks and mIoU$_{inv}$ between the occluded region to quantify the ability of reasoning objects with partial information. Higher values in both metrics represent better performance.

**Implementation Details**. We conducted all experiments on the PyTorch platform by using an NVIDIA RTX 4090 GPU. Our CondDiff-AMO employed pretrained PVTv2-B4 in the PCN for initialization, and all images are set to $256 \times 256$ same as all baselines we compared with in experiments. We utilized AdamW (Loshchilov and Hutter 2017) as the optimizer and set the batch size to 16. The learning rate applies to a cosine adjustment strategy, starting at $0.0001$ across $40$ epochs. During the sampling period, the inferring step is set to $T = 10$ for all experiments.

## 4.2 Comparisons with State-of-the-Art Methods

**Baselines.** To better evaluate our proposed model, we selected several typical and recent approaches for comparison, including C2F-Seg (Gao et al. 2023), HORI and GMC (Zhang et al. 2024), OccAmodal and SDAmodal (Zhan et al. 2024) for the supervised experiments; C2F-Seg (retrain), SD-XL Inpainting (Podell et al. 2023), Pix2gestalt (Ozguroglu et al. 2024) and DiT-XL (Ravishankar et al. 2025) for zero-shot learning. C2F-Seg, GMC and OccAmodal consist of two training stages, while SDAmodal and Pix2gestalt contains one-stage training (the same as our CondDiff-AMO). We reproduced C2F-Seg results by training its official scripts, denoted as C2F-Seg (retrain).

| **Method** | **Venue** | **COCOA** | |
|---|---|---|---|
| Supervised Exp. | | mIoU$_{\text{full}}$↑ | mIoU$_{\text{inv}}$↑ |
| Deocclusion_One | CVPR20 | 88.0 | 63.8 |
| Deocclusion_Two | CVPR20 | 88.2 | 65.3 |
| ASBU | ICCV21 | 88.9 | 65.3 |
| C2F-Seg (retrain) | ICCV23 | 80.5 | 36.5 |
| HORI | AAAI24 | 86.4 | - |
| GMC | AAAI24 | 86.9 | - |
| SDAmodal | CVPR24 | 90.7 | 71.6 |
| **Ours** | | **90.8** | **73.3** |

Table 1: Amodal segmentation results on COCOA datasets. Models were trained and evaluated on COCOA.

| **Method** | **Venue** | **COCOA** | **MP3D** |
|---|---|---|---|
| Zero-Shot Exp. | | mIoU$_{\text{full}}$ ↑ | mIoU$_{\text{full}}$ ↑ |
| C2F-Seg (retrain) | ICCV23 | 78.7 | 68.9 |
| SD-XL Inpainting | ICLR24 | 76.5 | - |
| Pix2gestalt | CVPR24 | 82.9 | 61.5 |
| DiT-XL | CVPR25 | 82.9 | 63.9 |
| **Ours** | | **83.0** | **71.6** |

Table 2: Zero-shot amodal segmentation results on COCOA and MP3D datasets.

Pix2gestalt and DiT-XL are state-of-the-art methods designed for zero-shot amodal segmentation tasks. SD-XL Inpainting and SAM are designed for general image generation and segmentation tasks. More methods (Xiao et al. 2021; Zhan et al. 2020; Nguyen and Todorovic 2021; Tran et al. 2022) were included in the supervised experiments. Deocclusion has both one-stage and two-stage architectures, so we followed SDAmodal to denote them as Deocclusion_One and Deocclusion_Two.

**Results**. As shown in Tables 1, 2 and 3, we organized two kinds of experiments, supervised experiments and zero-shot experiments. For the supervised settings, models were trained and tested via the same datasets, whereas models were trained on the Pix2gestalt dataset and subsequently evaluated using the other five benchmarks for zero-shot experiments. Our CondDiff-AMO significantly surpassed existing methods and achieved state-of-the-art results on all five datasets in terms of the two settings.Under supervised learning, CondDiff-AMO attained outstanding 90.8% mIoU$_{full}$ and 73.3% mIoU$_{inv}$ in COCOA. On the KINS, COCOA-cls and D2SA, our model obtained at least 7.0% and 3.0% improvement under supervised and zero-shot learning, respectively, compared to the second-best method. Several qualitative results under two settings are also shown in Fig. 5. Our method showed our superiority by providing more accurate and complete masks.

## 4.3 Ablation Studies

**The Effect of Model Design.** Our model design included multiple proposed modules, which were ablated one-by-one in the upper part of Table 4. By simply using images as input

| Method | Venue | KINS | | COCOA-cls | | D2SA | |
|---|---|---|---|---|---|---|---|
| Supervised Exp. | | $\text{mIoU}_{\text{full}}\uparrow$ | $\text{mIoU}_{\text{inv}}\uparrow$ | $\text{mIoU}_{\text{full}}\uparrow$ | $\text{mIoU}_{\text{inv}}\uparrow$ | $\text{mIoU}_{\text{full}}\uparrow$ | $\text{mIoU}_{\text{inv}}\uparrow$ |
| Deocclusion_One | CVPR20 | 78.02 | 38.14 | 76.91 | 20.34 | 80.45 | 28.56 |
| AISformer | BMVC22 | 81.53 | 48.54 | 72.69 | 13.75 | 86.81 | 30.01 |
| VRSP | AAAI21 | 80.70 | 47.33 | 78.98 | 22.92 | 88.08 | 35.17 |
| C2F-Seg | ICCV23 | 82.22 | 53.60 | 80.28 | 27.71 | 89.10 | 42.72 |
| **Ours** | - | **95.21** | **77.73** | **87.92** | **55.08** | **96.85** | **80.56** |
| Zero-Shot Exp. | | | | | | | |
| SAM | ICCV23 | 75.88 | - | 73.10 | - | 84.65 | - |
| C2F-Seg (retrain) | ICCV23 | 70.14 | - | 75.44 | - | 83.89 | - |
| SDXL-Inpainting | ICLR24 | 76.19 | - | 73.65 | - | 80.53 | - |
| Pix2gestalt | CVPR24 | 81.45 | - | 79.08 | - | 81.82 | - |
| **Ours** | - | **85.07** | - | **82.57** | - | **87.52** | - |

Table 3: Amodal segmentation results on five benchmarks. The results of all supervised methods were reported from C2F-Seg and zero-shot performance was recorded from Pix2gestalt, except for C2F-Seg results (obtained by retraining its official scripts). Note that '-' represents the missing data in Pix2gestalt.

| ID | Model Design | | | LS | | COCOA | |
|---|---|---|---|---|---|---|---|
| | VM | OCR | AFG | UST | VGPE | $\text{mIoU}_{\text{full}}\uparrow$ | $\text{mIoU}_{\text{inv}}\uparrow$ |
| A | × | × | × | | | 76.5 | 45.5 |
| B | × | | | | | 85.6 | 57.9 |
| C | | × | | | | 90.2 | 72.2 |
| D | | | × | | | 86.9 | 58.6 |
| E | × | × | | | | 85.1 | 58.4 |
| F | | × | × | | | 86.2 | 59.7 |
| G | × | | × | | | 77.9 | 49.8 |
| H | | | | × | × | 80.8 | 37.0 |
| I | | | | × | | 81.7 | 39.9 |
| J | | | | | × | 90.6 | 73.1 |
| **Ours** | | | | | | **90.8** | **73.3** |

Table 4: Ablation studies of model design and learning strategies. (×: removed, LS: learning strategies, VM: modal masks, OCR, AFG, UST, VGPE: refer to Section 3)

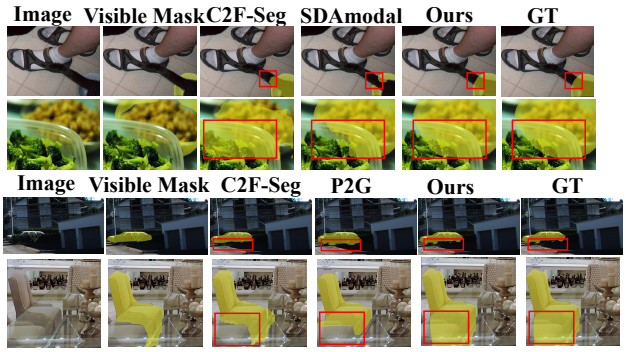


Figure 5: Qualitative comparison. The first two lines are from COCOA dataset (supervised learning) and the rest are from KINS and MP3D (zero-shot learning). The superior performance of the proposed method is enclosed by red rectangles. Zoom in for a better view. More results are in supplementary material.

without model masks (VM), two metrics showed the greatest degradation. The last-layer feature without OCR and its feature fusion, replaced AFG by LE, also illustrated a negative impact on two metrics.

**The Effect of Learning Strategy.** In the lower part of Table 4, we also evaluated the effectiveness of the UST and VGPE strategies applied in our proposed CondDiff-AMO. Specifically, models H and I caused obvious degradation to prove that the UST is essential for inferring hidden regions and the necessity to explore characteristics of amodal segmentation in diffusion models. Adding VGPE still showed slight improvement, based on model J and the proposed model.

## 5 Conclusion

In this paper, we designed a novel diffusion-based framework, CondDiff-AMO, for amodal segmentation in either supervised or zero-shot configurations, leveraging a diffusion algorithm to aid amodal mask prediction. Due to the uniqueness of amodal segmentation, we introduced three components, an UST training strategy for intimating inferring procedure from modal masks to amodal masks, PCN cooperating with DN network to produce adaptively conditional priors from images and masks over time, and a VGPE sampling strategy to restrict the prediction process (adapting well to the benefits of the diffusion algorithm to further enhance model's performance). Experimental results demonstrate that CondDiff-AMO significantly outperforms existing zero-shot and even supervised approaches achieving competitive scores over five benchmarks. The ablation studies also confirm the effectiveness of each component in CondDiff-AMO. As part of future work, we will extend this framework to amodal appearance completion, which involves synthesizing RGB pixel values within masks.

## Acknowledgments

This work was supported by the Science and Technology Development Fund, Macao S.A.R (FDCT) (Grant no. 0028/2023/RIA1).

## References

Amit, T.; Shaharbany, T.; Nachmani, E.; and Wolf, L. 2021. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*.

Baranchuk, D.; Rubachev, I.; Voynov, A.; Khrulkov, V.; and Babenko, A. 2021. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*.

Breitenstein, J.; and Fingscheidt, T. 2022. Amodal cityscapes: a new dataset, its generation, and an amodal semantic segmentation challenge baseline. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, 1018–1025. IEEE.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848.

Chen, Z.; Sun, K.; and Lin, X. 2024. CamoDiffusion: Camouflaged object detection via conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2, 1272–1280.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Ehsani, K.; Mottaghi, R.; and Farhadi, A. 2018. Segan: Segmenting and generating the invisible. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 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.

Gao, J.; Qian, X.; Wang, Y.; Xiao, T.; He, T.; Zhang, Z.; and Fu, Y. 2023. Coarse-to-fine amodal segmentation with shape prior. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1262–1271.

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. *Advances in neural information processing systems*, 33: 6840–6851.

Inagaki, Y.; Araki, R.; Yamashita, T.; and Fujiyoshi, H. 2019. Detecting layered structures of partially occluded objects for bin picking. In 2019 IEEE. In *RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5786–5791.

Kanazawa, A.; Tulsiani, S.; Efros, A. A.; and Malik, J. 2018. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European conference on computer vision (ECCV)*, 371–386.

Kar, A.; Tulsiani, S.; Carreira, J.; and Malik, J. 2015. Category-specific object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1966–1974.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Lee, S.; Chung, H.; Kim, J.; and Ye, J. C. 2022. Progressive deblurring of diffusion models for coarse-to-fine image synthesis. *arXiv preprint arXiv:2207.11192*.

Li, K.; and Malik, J. 2016. Amodal instance segmentation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 677–693. Springer.

Li, Z.; Ye, W.; Jiang, T.; and Huang, T. 2022. 2D amodal instance segmentation guided by 3D shape prior. In *European Conference on Computer Vision*, 165–181. Springer.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Nguyen, K.; and Todorovic, S. 2021. A weakly supervised amodal segmenter with boundary uncertainty estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7396–7405.

Ozguroglu, E.; Liu, R.; Surís, D.; Chen, D.; Dave, A.; Tokmakov, P.; and Vondrick, C. 2024. pix2gestalt: Amodal segmentation by synthesizing wholes. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3931–3940. IEEE Computer Society.

Palmer, S. E. 1999. *Vision science: Photons to phenomenology*. MIT press.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Qi, L.; Jiang, L.; Liu, S.; Shen, X.; and Jia, J. 2019. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3014–3023.

Ravishankar, R.; Patel, Z.; Rajasegaran, J.; and Malik, J. 2025. Scaling properties of diffusion models for perceptual tasks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12945–12954.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Sun, K.; Chen, Z.; Lin, X.; Sun, X.; Liu, H.; and Ji, R. 2025. Conditional diffusion models for camouflaged and salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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*.

Wada, K.; Kitagawa, S.; Okada, K.; and Inaba, M. 2018. Instance segmentation of visible and occluded regions for finding and picking target from a pile of objects. In *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 2048–2055. IEEE.

Wada, K.; Okada, K.; and Inaba, M. 2019. Joint learning of instance and semantic segmentation for robotic pick-and-place with heavy occlusions in clutter. In *2019 international conference on robotics and automation (ICRA)*, 9558–9564. IEEE.

Wang, J.; Huang, Q.; Tang, F.; Meng, J.; Su, J.; and Song, S. 2022a. Stepwise feature fusion: Local guides global. In *International conference on medical image computing and computer-assisted intervention*, 110–120. Springer.

Wang, M.; Ding, H.; Liew, J. H.; Liu, J.; Zhao, Y.; and Wei, Y. 2023. SegRefiner: Towards model-agnostic segmentation refinement with discrete diffusion process. *arXiv preprint arXiv:2312.12425*.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022b. Pvt v2: Improved baselines with pyramid vision transformer. *Computational visual media*, 8(3): 415–424.

Whang, J.; Delbracio, M.; Talebi, H.; Saharia, C.; Dimakis, A. G.; and Milanfar, P. 2022. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16293–16303.

Wu, J.; Ji, W.; Fu, H.; Xu, M.; Jin, Y.; and Xu, Y. 2024. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, 6, 6030–6038.

Wu, S.; Li, R.; Jakab, T.; Rupprecht, C.; and Vedaldi, A. 2023. Magicpony: Learning articulated 3d animals in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8792–8802.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, 4, 2995–3003.

Yao, J.; Hong, Y.; Wang, C.; Xiao, T.; He, T.; Locatello, F.; Wipf, D. P.; Fu, Y.; and Zhang, Z. 2022. Self-supervised amodal video object segmentation. *Advances in neural information processing systems*, 35: 6278–6291.

Zhan, G.; Zheng, C.; Xie, W.; and Zisserman, A. 2024. Amodal ground truth and completion in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28003–28013.

Zhan, X.; Pan, X.; Dai, B.; Liu, Z.; Lin, D.; and Loy, C. C. 2020. Self-supervised scene de-occlusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3784–3792.

Zhang, B.; Liu, Q.; Zhang, J.; Wang, Y.; Liu, L.; Lin, Z.; and Liu, Y. 2024. Amodal scene analysis via holistic occlusion relation inference and generative mask completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7, 6997–7005.

Zhu, Y.; Tian, Y.; Metaxas, D.; and Dollár, P. 2017. Semantic amodal segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1464–1472.

Zou, C.; and Hoiem, D. 2020. Silhouette guided point cloud reconstruction beyond occlusion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 41–50.