Tuning-Free Amodal Segmentation via the Occlusion-Free Bias of Inpainting Models

Jae Joong Lee, Bedrich Benes, Raymond A. Yeh

Department of Computer Science, Purdue University
West Lafayette, IN 47906 USA
{lee2161, bbenes, rayyeh}@purdue.edu

Abstract

Amodal segmentation is an image-based algorithm that aims to predict masks for both visible and occluded parts of objects. Existing methods typically rely on supervised learning with annotated amodal masks or synthetic data. The effectiveness of these methods relies heavily on the quality of the datasets. This dependence can unintentionally restrict their generalization capabilities due to insufficient diversity and size. Although existing zero-shot methods perform well on their reported datasets, their performance does not necessarily transfer to other datasets. We propose a **tuning-free** approach that re-purposes diffusion-based inpainting foundation models for amodal segmentation. Our approach is motivated by the "occlusion-free bias" of inpainting models, i.e., the inpainted objects tend to be complete and without occlusions. We reconstruct the occluded regions of an object via inpainting and then apply segmentation, all **without additional training or fine-tuning.** Experiments on five datasets, three previously unreported, demonstrate the generalizability of our approach. On average, our approach achieves 5.3% more accurate masks in mIoU compared to the publicly available state-of-the-art, pix2gestalt.

1 Introduction

Amodal segmentation refers to predicting segmentation masks even under occlusions (Li and Malik 2016). This challenging task involves reasoning about the unseen portion of an object under complex occlusion and illumination scenarios. It is an important problem with potential applications in autonomous driving and robot planning, which require reasoning beyond what is directly observed to predict possible future events in the environment (Yang et al. 2019; Geiger, Lenz, and Urtasun 2012; Dang et al. 2019).

Following the success of deep segmentation methods (Kirillov et al. 2023; Caron et al. 2021; Long, Shelhamer, and Darrell 2015; Ronneberger, Fischer, and Brox 2015), amodal segmentation is often formulated as a supervised learning task, *i.e.*, a dataset of (image, amodal mask) pairs is collected to train a model. However, preparing a large dataset for amodal segmentation is challenging, as annotating amodal masks requires reasoning over occluded regions, which may be difficult and inconsistent among human annotators. Furthermore, scaling the diversity of the dataset requires numerous combinations of occluders and objects. Inevitably, several amodal segmentation methods turns to synthetically generating occlusions (Xiao et al. 2021; Follmann et al. 2019; Ozguroglu et al. 2024; Ao, Ke, and Ehinger 2024) and 3D game engine rendering (Hu et al. 2019) to obtain the annotations. However, the performance is still limited by (a) the distribution gap between the synthetic and real data and (b) the size of the dataset. For example, the currently available state-of-the-art (SOTA) (Ozguroglu et al. 2024) uses only 800k data pairs for training/fine-tuning, which is relatively small compared to the recent Internet-scale datasets for other tasks (Rombach et al. 2022; Peebles and Xie 2023; Podell et al. 2023). Another work (Xu, Zhang, and Shi 2024) also proposes using a pre-trained diffusion model, but it requires iterative occlusion removal and is constrained to a *fixed 83 object categories*, which has limited generalizability.

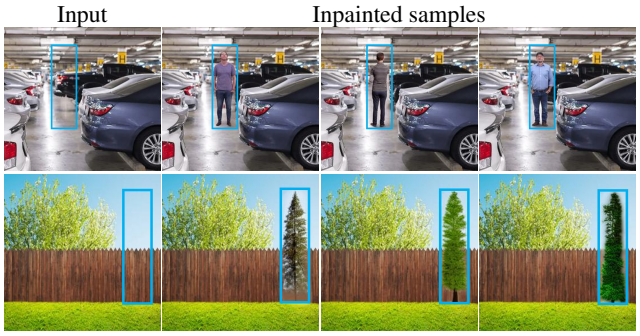

Figure 1: Occlusion-free bias for a diffusion inpainting model. We observe that an inpainted object is always placed without occlusions inside the inpainting area (blue box), *e.g.*, a tree could have been inpainted behind the fence.

To address these challenges, we present a *tuning-free approach* that utilizes existing foundation models trained on Internet-scale datasets. Our method **does not require** any amodal data or any training data at all. Hence, the method is naturally zero-shot and without restriction to pre-defined object classes.

The approach is motivated by the observation that diffusion inpainting models have an "occlusion-free bias", *i.e.*, the inpainting model prefers to generate a whole object rather than the occluder given a reasonable mask as shown

in Fig. 1. We propose to perform inpainting over an enlarged modal mask, where the diffusion model fills the occluded regions. With the inpainted occluded regions, we extract the modal segmentation as the amodal prediction.

We demonstrate the effectiveness of our approach on five diverse amodal segmentation datasets: COCO-A (Zhu et al. 2017a), BSDS-A (Zhu et al. 2017b), KINS (Qi et al. 2019), FishBowl (Tangemann et al. 2021), and SAILVOS (Hu et al. 2019). Our zero-shot and tuning-free approach outperforms the *supervised* current SOTA (Ozguroglu et al. 2024) by 5.3% in all datasets on average in the mIoU, and a more notable 12.1% gain from previously unreported three datasets.

Our contributions are as follows:

- We propose a tuning-free method for amodal segmentation (zero-shot) by exploiting the occlusion-free bias of diffusion inpainting models.
- The method involves several novel components, including a context-aware approach to background composition using RGB distribution, a noising process image for conditioning, and a modal mask construction procedure.
- We demonstrate the generalizability of the proposed method by conducting extensive experiments over four diffusion inpainting models on five diverse datasets.

2 Related Work

Amodal perception and segmentation. Humans can often detect and identify an object even if it is (partially) occluded (Lehar 1999). Seminal work by Li and Malik (2016) begins the line of work of using deep learning for amodal tasks. Many architectures and models have been proposed, *e.g.*, CNN (Li and Malik 2016; Zhang et al. 2019; Yang et al. 2019; Xiao et al. 2021), Generative Adversarial Networks (Ehsani, Mottaghi, and Farhadi 2018), Transformer (Tran et al. 2022; Gao et al. 2023), and Diffusion-based models (Ozguroglu et al. 2024; Zhan et al. 2024; Chen, Ramanan, and Khurana 2025). These works are evaluated on datasets such as COCO-A (Zhu et al. 2017a), BSDS-A (Zhu et al. 2017b), KINS (Qi et al. 2019), and MP3D-Amodal (Zhan et al. 2024), which consist of common objects from the real world. Other synthetic benchmarks are also popular, *e.g.*, SAILVOS (Hu et al. 2019) or FishBowl (Tangemann et al. 2021). These synthetic datasets provide more diverse object categories and precise amodal mask annotations without human errors.

More recently, SegmentAnything-based models (Liu et al. 2025; Tai et al. 2025) have been proposed by training on amodal datasets; however, the code is not available. The current SOTA (with public code) are pix2gestalt (Ozguroglu et al. 2024) and AmodalWild (Zhan et al. 2024). We refer to pix2gestalt as the SOTA as its code is fully released and reproducible.

The pix2gestalt (Ozguroglu et al. 2024) trains a deepnet to predict the occluded pixels following an analysis by synthesis framework (Yuille and Kersten 2006). A synthetic amodal dataset is created by occluding objects with randomly sampled overlays using another object. Our work does not require any amodal datasets, *i.e.*, it is tuning-free.

The closest related to our work is the tuning-free method proposed by Xu, Zhang, and Shi (2024), which proposes to iteratively use an inpainting modal for amodal completion, *i.e.*, they are interested in high image quality. Nonetheless, as a modal mask can be extracted from the completed image, it is considered an amodal segmentation method.

Finally, a key limitation of Xu, Zhang, and Shi (2024) and Zhan et al. (2024), upon reviewing their code, is that these approaches leverage pre-defined class information, *e.g.*, Xu, Zhang, and Shi (2024) is limited to only 83 classes[1], which greatly limits the usability. In contrast, our method does not have class restrictions and does not require multiple calls to the inpainting model.

Image inpainting is the task of filling in missing regions of a given image, where the missing region is indicated using a mask. Early works in inpainting leverage low-level properties of natural images, *e.g.*, smoothness (Tschumperlé and Deriche 2005; Darabi et al. 2012) or low-rank (Jin and Ye 2015; Guo et al. 2017), to tackle this task. In cases where the image contains a large missing region, then generative or learning-based methods are proposed (Yeh et al. 2017; Yu et al. 2019; Zeng et al. 2020; Li et al. 2020; Guo, Yang, and Huang 2021; Li et al. 2022; Liu et al. 2022; Yildirim et al. 2023). More recently, diffusion models have emerged as the state-of-the-art in image generation and image inpainting methods based on diffusion have also been proposed (Lugmayr et al. 2022; Suvorov et al. 2022; Rombach et al. 2022; Saharia et al. 2022; Liu et al. 2024; Corneanu, Gadde, and Martinez 2024). Differently, this work leverages pre-trained diffusion inpainting models for the task of zero-shot amodal segmentation.

Tuning-free methods for diffusion models. A diffusion model requires training on a large number of images to generate a realistic output. Although pre-trained diffusion models exist, fine-tuning is still needed for new tasks. Recent tuning-free methods leverage pre-trained models for performance gains and other tasks without extra training, improving Text-to-Image synthesis (Zeng et al. 2024; Ding et al. 2024) and video generation (Qiu et al. 2024; He et al. 2023). Similar to these works, our approach does not require additional tuning. Differently, this paper focuses on using pre-trained inpainting models for the *task of amodal segmentation*.

3 Preliminaries

Diffusion models add noise to the data (forward process) and learn to undo the added noise (reverse process) during training. For generations, diffusion models start from a purely sampled noise and perform the reverse process. **Forward diffusion process** gradually adds Gaussian noise to a clean image, $\boldsymbol{x}_0$, over T timesteps where $\boldsymbol{x}_t$ is the noisy version of the image at timestep t with α_t controlling the amount of noise added at each step. The noisy image x_t can be computed from x_0 as follows

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{1}$$

[1] github.com/k8xu/amodal/blob/main/classes.txt

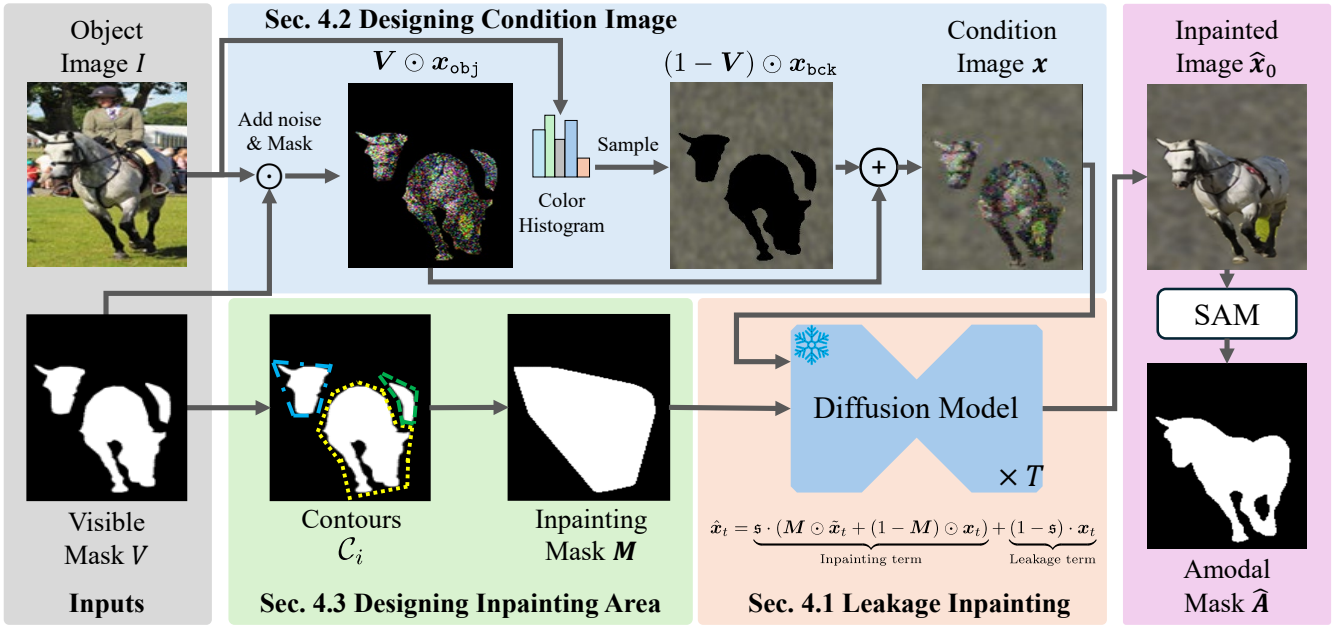

Figure 2: Our approach takes two inputs: an RGB image $\boldsymbol{I}$ and a visible mask $\boldsymbol{V}$. From $\boldsymbol{I}$, we generate a conditioned RGB image with a color distribution-aware background $\boldsymbol{x}_{\texttt{bck}}$ and a partial Gaussian noise-added object $\boldsymbol{x}_{\texttt{obj}}$. From $\boldsymbol{V}$, we create a customized inpainting area $\boldsymbol{M}$ so that we utilize any diffusion-based inpainting models to create an inpainted image $\hat{\boldsymbol{x}}_0$ to extract an amodal mask $\hat{\boldsymbol{A}}$.

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ is the cumulative product of the noise scaling factors and $\epsilon \sim \mathcal{N}(0, I)$ is a Gaussian noise.

Reverse diffusion process undoes the forward diffusion by denoising an image iteratively, starting from a pure noise image, $\boldsymbol{x}_T$ to the clean image, $\boldsymbol{x}_0$. This is formulated as a sequence of conditional probabilities

$$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \sigma_t \boldsymbol{I}) \tag{2}$$

following the Gaussian distribution with mean $\mu_\theta(\boldsymbol{x}_t, t)$, and diagonal covariance matrix predicted from a deep-net with parameters θ. Intuitively, μ_θ can be thought of as an image denoiser that gradually removes noise according to a schedule. Another choice, introduced by DDPM (Ho, Jain, and Abbeel 2020), is to use a deep-net to model the residual noise ϵ_θ. This is equivalent to choosing a denoiser

$$\mu(\boldsymbol{x}, t) \triangleq \boldsymbol{x} - \sigma_t \cdot \epsilon_\theta(\boldsymbol{x}, t). \tag{3}$$

Inpainting predicts masked-out regions of a given input image. Diffusion-based method (Lugmayr et al. 2022) leverages the generative prior of a pre-trained DDPM (Ho, Jain, and Abbeel 2020) to do so. This is achieved by iteratively removing noise from the linear combination of the noisy unmasked regions with the generated mask regions. More formally, given an input image $\boldsymbol{x}$ and a mask $\boldsymbol{M} \in \{0,1\}^{H \times W}$ the generation process to produce an inpainted image $\hat{\boldsymbol{x}}_0$ is as follows:

$$\tilde{\boldsymbol{x}}_t \sim \mathcal{N}\left(\mu_\theta(\hat{\boldsymbol{x}}_{t-1}, t), \sigma_{t-1}\right) \tag{4}$$

$$\hat{\boldsymbol{x}}_t = \boldsymbol{M} \odot \tilde{\boldsymbol{x}}_t + (1 - \boldsymbol{M}) \odot \boldsymbol{x}_t, \tag{5}$$

where $\boldsymbol{x}_t$ is the noise added input image following Eq. (1), $\tilde{\boldsymbol{x}}_T$ is assumed to be pure noise, and $\odot$ denotes element-wise multiplication. Recent foundation diffusion models (Rombach et al. 2022; Podell et al. 2023; Labs 2024) are text-conditioned. The denoiser takes in an additional text prompt $\boldsymbol{c}$ to guide the generation, *i.e.*,

$$\tilde{\boldsymbol{x}}_t(\boldsymbol{c}) \sim \mu_\theta\left(\hat{\boldsymbol{x}}_{t-1}, \boldsymbol{c}, t\right). \tag{6}$$

4 Training-free Amodal Segmentation

Problem formulation. We consider the amodal segmentation setup as in Ozguroglu et al. (2024). Given an object's image $\boldsymbol{I}$ and corresponding visible (modal) mask $\boldsymbol{V}$, the task is to predict the object's amodal mask $\hat{\boldsymbol{A}}$ that covers the whole object, including occluded regions.

Overview. We propose a tuning-free method for amodal segmentation by re-purposing diffusion inpainting models. Our approach leverages the "occlusion-free" bias of diffusion inpainting models, as shown in Fig. 1, where an inpainted object is almost always generated without occlusion. Hence, we inpaint an occluded object to remove the occlusion and use a segmentation method, *e.g.*, SAM (Kirillov et al. 2023), on the unoccluded object to extract the amodal mask $\hat{\boldsymbol{A}}$. While the proposed method seems straightforward, the devil is in the details.

To achieve high-quality amodal masks, we needed to carefully design the generation procedure of the inpainting model (Sec. 4.1), the conditioning image $\boldsymbol{x}$ (Sec. 4.2), and the inpainting area $\boldsymbol{M}$ (Sec. 4.3). A visual illustration of the approach is provided in Fig. 2.

Figure 3: Our soft-inpainting approach is based on diffusion that can extrapolate to an amodal mask much larger than the visible mask, as shown in this sequence of denoising images.

4.1 Inpainting via leakage conditioning

Recent diffusion inpainting models (Rombach et al. 2022; Podell et al. 2023; Labs 2024) are often text-conditioned, *i.e.*, the model performs a conditional generation on the masked area with a text-prompt.

As the task of amodal segmentation does not involve any text prompt, we need another way to condition the model. Specifically, besides the standard diffusion sampling for inpainting, we further "leak" the original unmasked conditioning image $\boldsymbol{x}$ to the model. Instead of Eq. (5), we perform the following:

$$\hat{\boldsymbol{x}}_t = \underbrace{\mathfrak{s} \cdot (\boldsymbol{M} \odot \tilde{\boldsymbol{x}}_t + (1 - \boldsymbol{M}) \odot \boldsymbol{x}_t)}_{\text{Inpainting term}} + \underbrace{(1 - \mathfrak{s}) \cdot \boldsymbol{x}_t}_{\text{Leakage term}}, \quad (7)$$

where $\mathfrak{s} \in \mathbb{R}^+$ controls the strength of the leakage. The purpose of the leakage term is to inform the model to inpaint occluded parts relevant to the current scene context, both masked and non-masked regions. Empirically, we set $\mathfrak{s} = 0.3$ to balance the level of image context preservation, which is equivalent to using an image that is a combination of 30% of a newly generated image and 70% of the original image to maintain the overall original context. Increasing noise loses visual contexts, leading to random objects with poor quality of amodal segmentation.

The update equation in Eq. (7) *no longer strictly* performs image inpainting as the non-masked region is *not* guaranteed to be the same as the conditioning image $\boldsymbol{x}$. Instead, we perform a "soft"-inpainting where the model generates an image that roughly resembles the condition image $\boldsymbol{x}$ for the unmasked regions and focuses on generating within the inpainting area. As the inpainting area is not strict, this also has the benefit that **pixels outside of the inpainting area $\boldsymbol{M}$ can be changed**, *i.e.*, the predicted amodal mask can be larger than the inpainting area $\boldsymbol{M}$, which helps with cases where extrapolation of the visible mask is needed. To handle this, we leverage the leakage from eq. (7), which acts like a "soft scaffold", allowing the model to perform "soft-inpainting", changing pixels outside the mask (see fig. 3 and fig. 8).

4.2 Designing the condition image

The inpainting procedure needs an input condition image, denoted as $\boldsymbol{x}$, to guide the generation process. Our objective is to create a complete object without any occlusions. To achieve this, we want the model to focus on the visible parts of the object rather than the background. Therefore, we have a separate procedure for preparing the object and background pixels, where

$$\boldsymbol{x} = \boldsymbol{V} \odot \boldsymbol{x}_{\texttt{obj}} + (1 - \boldsymbol{V}) \odot \boldsymbol{x}_{\texttt{bck}}. \quad (8)$$

Object pixels. Given the image $\boldsymbol{I}$ containing the object and its corresponding visible mask $\boldsymbol{V}$, we extract the object pixels by overlaying the visible mask using an element-wise multiplication. As a diffusion model expects a noisy image, we add noise to the object pixels similar to (Corneanu, Gadde, and Martinez 2024; Meng et al. 2022), *i.e.*,

$$\boldsymbol{x}_{\texttt{obj}} = (\mathfrak{s} \cdot \epsilon + (1 - \mathfrak{s}) \cdot \boldsymbol{I}), \quad (9)$$

where $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$ and $\mathfrak{s} = 0.3$.

Background pixels. In standard inpainting, the pixel values in the masked-out region (background) are irrelevant. Hence, it is common to choose either black or white. However, the background now plays a role due to our leakage conditioning in Eq. (7). The default choice of black or white introduces a sharp contrast around the object's contours, and diffusion models do not react well to this pixel intensity discontinuity.

Inspired by previous works to blend images seamlessly, such as leveraging the denoising process (Lugmayr et al. 2022) and incorporating latent information from text-guided diffusion models (Avrahami, Lischinski, and Fried 2022). We construct a smooth background that matches the object's color distribution. First, we build a color histogram from the object's visible pixels in $\boldsymbol{I}$, then sample background pixels $\boldsymbol{x}_{\texttt{bck}}$ based on histogram frequencies, and finally apply a Gaussian blur.

4.3 Designing the soft inpainting area

Besides the condition image $\boldsymbol{x}$, the inpainting procedure also requires an inpainting area $\boldsymbol{M}$, which specifies where to focus on the generation.

From the visible mask $\boldsymbol{V}$, we extract contours from a set of points corresponding to visible regions, where $\mathcal{C}_i$ is the i^{th} contour. Next, we combine the contours into one region by taking their union and finding the smallest convex polygon $\texttt{CnvxHull}(\bigcup_{i=1} \mathcal{C}_i)$ that can enclose all contours. Finally, we get the inpainting region, $\boldsymbol{M}$, by setting values inside the convex polygon to one, where a visible pixel at (x, y):

$$\boldsymbol{M} = \begin{cases} 1, & \text{if } (x, y) \in \texttt{CnvxHull}(\bigcup_{i=1} \mathcal{C}_i), \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

Diffusion inpainting models are then trained to take an inpainting region as conditioning input, *i.e.*, $\tilde{\boldsymbol{x}}_t(\boldsymbol{M})$. Hence, this allows us to use classifier-free guidance (Ho and Salimans 2022) with the mask during the generation. Let w denote the intensity of the strictness, which is how the model

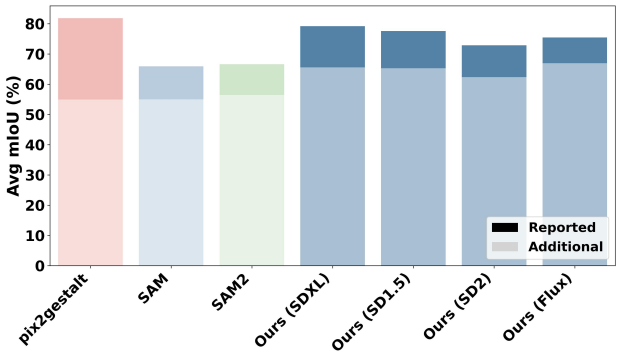

Figure 4: Stability analysis across five datasets. Pix2Gestalt, despite being **zero-shot**, shows a 26.9% drop in mIoU between its reported and unreported datasets, but our **tuning-free** method using Flux shows only an 8.5% performance difference. This shows that ours shows far more stability and generalizes unseen data better.

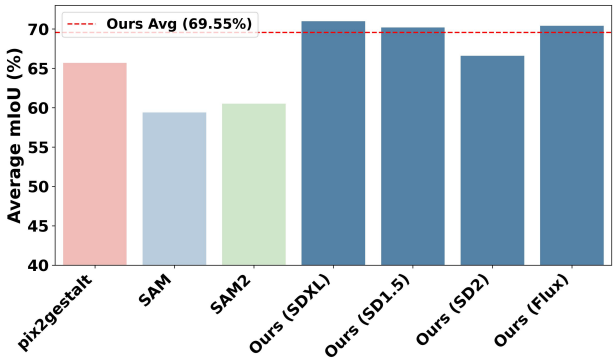

Figure 5: Average mIoU across five comprehensive datasets, showing our training-free approach using SDXL outperforms Pix2Gestalt (Ozguroglu et al. 2024), which needs training, by 5.3%.

needs to follow the conditioning of the inpainting region $\boldsymbol{M}$ a classifier-free-guided sample computes

$$\tilde{\boldsymbol{x}}_t^{\texttt{CFG}} = (1+w)\cdot\tilde{\boldsymbol{x}}_t(\boldsymbol{M}) - w\cdot\tilde{\boldsymbol{x}}_t(\varnothing), \tag{11}$$

where $\varnothing$ denotes the empty representation of $\boldsymbol{M}$. Instead of using $\tilde{\boldsymbol{x}}$ directly in Eq. (5), $\tilde{\boldsymbol{x}}_t^{\texttt{CFG}}$ is used. Intuitively, smaller w gives more freedom to generate new pixel information independent of the inpainting area shape. Empirically, we set $w = 0.75$ for Stable Diffusion version 1.5 (Rombach et al. 2022), Stable Diffusion XL (Podell et al. 2023) and w as 1.5 with Flux (Labs 2024).

5 Experiments

Our method is **tuning-free** and also a zero-shot amodal segmentation method. For a fair comparison, we strictly follow the experiment setup of the zero-shot amodal segmentation experiment setting by Ozguroglu et al. (2024) on COCO-A (Zhu et al. 2017a) and BSDS-A (Zhu et al. 2017b). To study the zero-shot capability, we evaluate three additional datasets, including KINS (Qi et al. 2019), FishBowl (Tangemann et al. 2021), and SAILVOS (Hu et al. 2019). We report quantitative and qualitative results, followed by ablations.

5.1 Experiment setup

We report on the following five datasets covering both real-world and synthetic images:

❶ COCO-A (Zhu et al. 2017a): Based on COCO (Lin et al. 2014) dataset, COCO-A (Zhu et al. 2017a) is a human-annotated amodal segmentation dataset over natural images. We report on its evaluation set with 13k ground truth object amodal annotations, including common objects.

❷ BSDS-A (Zhu et al. 2017b): Derived from the Berkeley Segmentation Dataset (BSDS) (Martin et al. 2001), BSDS-A (Zhu et al. 2017b) is an amodal segmentation dataset labeled with manual amodal annotation. We report on the evaluation image sets with 200 images from the real world.

❸ KINS (Qi et al. 2019): KINS (Qi et al. 2019), derived from KITTI (Geiger, Lenz, and Urtasun 2012) for autonomous driving, features manually annotated amodal masks and an evaluation set of 7k images.

❹ FishBowl (Fbowl) (Tangemann et al. 2021) is a synthetic dataset that has different numbers of fish from an WebGL demo (Greggman and Engines 2017). Its evaluation set contains 1k videos of 128 frames each, with each frame treated independently for amodal segmentation.

❺ SAILVOS (SV) (Hu et al. 2019) is a synthetic dataset from the photo-realistic game GTA-V. It contains 26k images along with 507k objects in the evaluation set.

Evaluation metric. Following Ozguroglu et al. (2024), we report the mean intersection over union (mIoU) to evaluate predicted amodal masks. A higher mIoU indicates a better match of the prediction with the ground truth. We also report the mIoU over different subsets of the data based on the occlusion rate of the object. Specifically, we report on occlusion rate subsets that are less than 50%. We observed that highly occluded objects yield uncertain annotations, as images often lack enough details for a complete amodal mask; See discussion in the appendix.

Baselines. We consider the state-of-the-art baseline of pix2gestalt, two training-required methods, and three additional tuning-free methods.

① pix2gestalt (Ozguroglu et al. 2024) takes an RGB image and its modal mask to generate an amodal mask by using SAM (Kirillov et al. 2023) to collect on a customized training dataset that has more than 800k image pairs with occlusions to a fine-tuned a *pre-trained diffusion model* of StableDiffusion2 (SD2) (Stability-AI 2022).

② Amodal Wild (Zhan et al. 2024) uses a two-stage approach. First, an occluder mask is predicted from an RGB image and its modal mask. Next, a U-Net-based model leveraging features from a pre-trained Stable Diffusion (using the modal mask and occluder boundary) predicts the amodal mask.

③ Inpaint-SDXL (Podell et al. 2023): Given a visible mask and an RGB image, SDXL inpaints its region by leveraging a pre-trained model to remove missing pixel information. This baseline is proposed by Ozguroglu et al. (2024) in

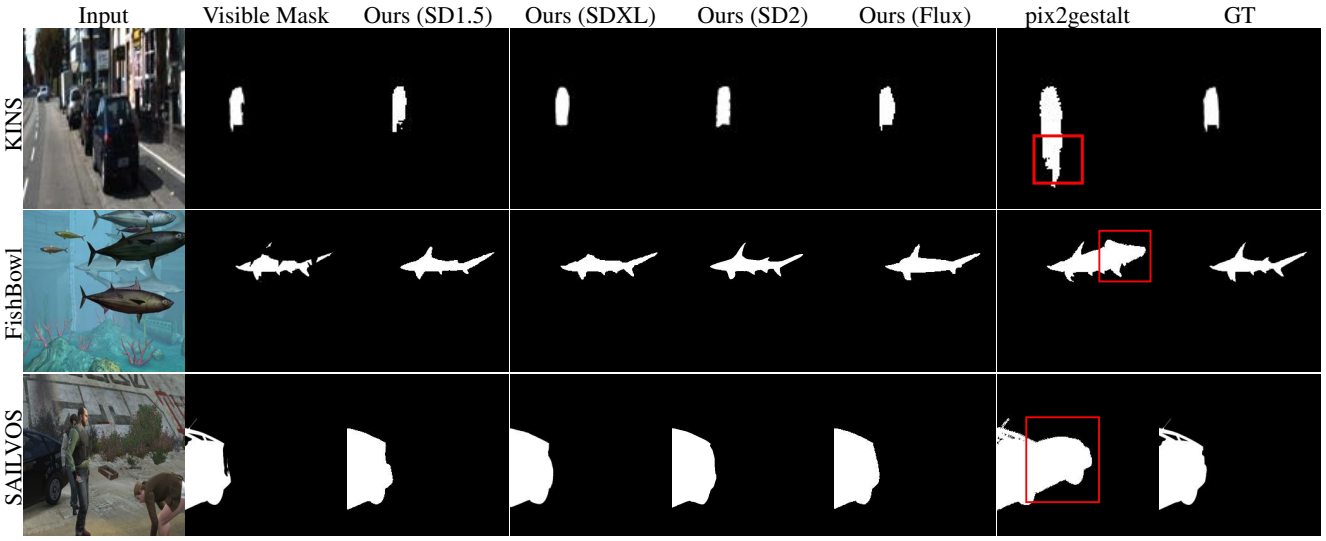

Figure 6: Qualitative comparison of amodal mask on KINS (Qi et al. 2019), FishBowl (Tangemann et al. 2021), and SAILVOS (Hu et al. 2019). We observe that for novel categories / domain pix2gestalt may hallucinate inaccurate amodal masks; see boxed regions in red.

pix2gestalt. We directly report the number from their paper, as the code for this baseline has not been released.

④ SAM (Kirillov et al. 2023): takes a set of points and an RGB image to segment pixels that fall into the same object category based on features from an image encoder. This is a strong modal baseline, as reported by Ozguroglu et al. (2024)

⑤ SAM2 (Ravi et al. 2024): We also consider a even strong modal baseline of SAM2, which is a improved version of SAM.

We also attempted to compare to Amodal Completion by Xu, Zhang, and Shi (2024). However, as it is limited to 83 categories, the approach was unable to generate a prediction for many images in the datasets we considered.

Implementation. We consider several diffusion models, including Stable Diffusion 1.5 and 2 (SD1.5, SD2) (Rombach et al. 2022), Stable Diffusion XL (SDXL) (Podell et al. 2023), and Flux (Labs 2024). Images are refined with 20 iterative steps for amodal completion, and the mask $\boldsymbol{M}$ is extracted by uniformly sampling nine points from $\boldsymbol{V}$ using SAM (Kirillov et al. 2023). All experiments were performed on an NVIDIA RTX 4090 (24GB VRAM), and 8-bit quantization was applied for Flux to reduce the memory usage.

5.2 Quantitative results

Recall, pix2gestalt **trains on a "synthetically curated dataset"** and is hence zero-shot. However, it is unclear whether this curated dataset generalizes beyond COCO-A and BSDS-A *e.g.*, if the testing distribution is very different from their curated data. In Fig. 4 we visualize the performance gap between pix2gestalt' reported datasets (COCO-A, BSDS-A) and three additional datasets. On one hand, there is a significant performance gap (26.9% difference) for pix2gestalt despite being a zero-shot method. On the other hand, our method has a gap of 8.5%. This illustrates the benefit of a tuning-free method that generalizes more effectively to samples outside the training distribution.

Method	Avg	COCO	BSDS	KINS	FBowl	SV	DiffMod
pix2gestalt	65.7	**82.9**	**80.8**	39.2	*73.3*	52.3	SD2
SAM	59.4	66.6	65.3	40.8	68.3	55.9	-
SAM2	60.5	70.1	63.1	46.9	65.5	57.0	-
Inpaint	-	76.5	74.2	-	-	-	SDXL
Ours	**71.0**	*82.7*	*75.6*	**60.4**	73.0	63.5	SDXL
Ours	70.2	79.9	75.2	58.4	71.0	**66.6**	SD1.5
Ours	66.6	73.2	72.6	57.2	72.8	57.2	SD2
Ours	*70.4*	75.5	75.5	*60.2*	**75.2**	*65.6*	Flux

Table 1: Quantiative comparisons of amodal mask in mIoU(%)↑. Methods, except for pix2gestalt are tuning-free. The best result is bolded, and the second best is *italicize*. The gray-colored columns show mIoU values from the unreported datasets, and ours have all the **best** performance.

Fig. 5 reports the average mIoU performance across the methods. Despite being tuning-free, Our **training-free**, SDXL-based approach achieves a 71.0% average mIoU versus 65.7% for pix2gestalt, representing a 5.3% improvement over pix2gestalt, which has been supervised on amodal data.

For further analysis, Tab. 1 reports mIoU using five datasets, and we **bold** the best metric and *italicize* the second-best metric. For the COCO-A (Zhu et al. 2017a) and BSDS-A (Zhu et al. 2017b), pix2gestalt (Ozguroglu et al. 2024) (a supervised method) has the best metrics, followed by Stable Diffusion XL with ours by 0.2% and 5.2%. Note, pix2gestalt "curated synthetic dataset" is designed to match the distribution of COCO-A and BSDS-A.

To further investigate the zero-shot capability, we present three additional datasets featuring diverse object categories. Results on KINS (Qi et al. 2019), FishBowl (FBowl) (Tange-

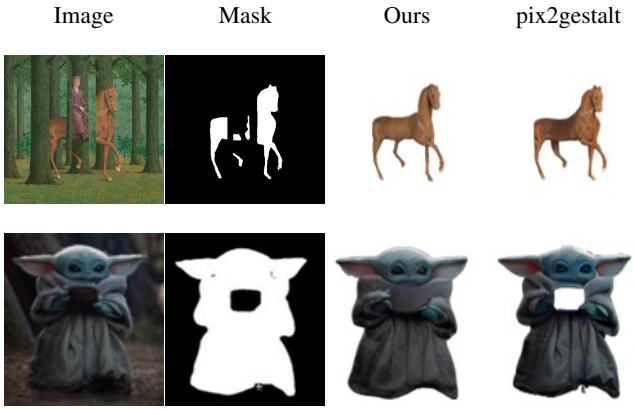

Figure 7: Amodal completion results on in-the-wild images comparing Ours (SDXL) and pix2gestalt.

mann et al. 2021), SAILVOS (SV) (Hu et al. 2019), show that our methods generate 21.2%, 1.9%, and 14.3% more accurate mask than pix2gestalt. Importantly, our method performs best among the tuning-free approaches and convincingly outperforms the modal baseline. We were unable to compare with Inpainting-SDXL, as the code was not released and we were unable to reproduce it.

In the appendix, we report and discuss the detailed results based on different object occlusion rates. Also, we provide further qualitative analysis for each dataset.

5.3 Qualitative results

Fig. 6 shows the comparison of the amodal mask generation on the three additional datasets, including KINS (Qi et al. 2019) (row 1), FishBowl (Tangemann et al. 2021) (row 2), and SAILVOS (Hu et al. 2019) (row 3). The first row shows that pix2gestalt overextends the car. The second row demonstrates that pix2gestalt misunderstands the visual context and adds "hallucinations", (another car), to generate amodal masks. Overall, pix2gestalt performs worse on these additional datasets, possibly due to a larger gap in distribution from their curated data. In contrast, our method shows robustness with high-quality amodal masks.

We also experimented with in-the-wild images to validate our approaches compared to pix2gestalt (Fig. 7). We start with the horse (first row) that pix2gestalt reported in their paper. The predicted mask from our approach shows a comparable quality to that generated by pix2gestalt. We also show another example (second row), where ours completes a cloth behind the cup that Grogu is holding, while pix2gestalt barely made any changes to the input image. Moreover, Fig. 8 shows cases of amodal mask extrapolation, demonstrating our method's ability to "grow" a bigger mask.

5.4 Ablation studies & analysis

Ablations. We evaluate the effectiveness of the components by removing them from the algorithm, and we show the results in Tab. 2. The experiment is conducted on COCO-A (Zhu et al. 2017a) using Stable Diffusion XL (Podell et al. 2023). The first row shows the mIoU with all the components included. When the leaking conditioning (Sec. 4.1), the context-aware background (Sec. 4.2), or the inpainting area (Sec. 4.3) is excluded, the accuracy of the amodal mask falls by 37.6%, 5.6%, 6.2% in the mIoU compared to using all components, respectively and the greater drop indicates higher importance for generating accurate amodal masks. We report additional ablation studies in the Appendix.

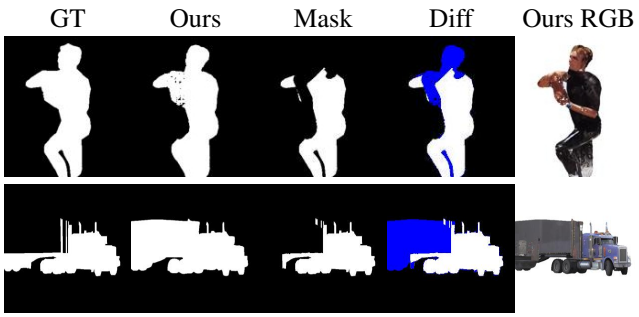

Figure 8: Blue pixels denote differences between the visible mask and our predicted mask. Observe that our method can extrapolate the mask via our proposed "soft-inpainting" method.

DiffMod	Leakage	Background	Mask	mIoU(%)↑
SDXL	✗	✓	✓	38.9
SDXL	✓	✗	✓	70.9
SDXL	✓	✓	✗	70.3
SDXL	✓	✓	✓	**76.5**

Table 2: We ablate each component's effectiveness using mIoU by removing each component from the pipeline.

Computation efficiency. We compare the efficiency to pix2gestalt. For the smallest model, SD2 is 4.1× more efficient in memory, and the inference (0.3 seconds) is 19× faster compared to pix2gestalt (Ozguroglu et al. 2024). Our best model (SDXL) is also more efficient than pix2gestalt.

6 Conclusion

We introduce a tuning-free/zero-shot amodal segmentation method by leveraging the occlusion-free bias of pre-trained diffusion inpainting models. Our approach customizes the conditioning image, designs a new inpainting region, and uses a novel leakage conditioning technique. Experiments on five datasets demonstrate that our model (SDXL) improves mIoU by **5.3%**, with **4.8**× faster inference and **1.4**× VRAM efficiency over pix2gestalt. Other models (SD1.5, SD2, Flux) are also effective. These results demonstrate the generalizability of our approach over existing supervised trained zero-shot methods. Finally, as diffusion inpainting techniques continue to improve, we anticipate further advancements in segmentation performance.

Acknowledgments. Work supported in part by: NSF 2506783, 2417510, 2412928, 2309564, USDA-NIFA 032382, 1032672, and a Google Research Scholar for RAY.

References

Ao, J.; Ke, Q.; and Ehinger, K. A. 2024. Amodal Intra-Class Instance Segmentation: Synthetic Datasets and Benchmark. In *WACV*.

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *CVPR*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *ICCV*.

Chen, K.; Ramanan, D.; and Khurana, T. 2025. Using Diffusion Priors for Video Amodal Segmentation. In *CVPR*.

Corneanu, C.; Gadde, R.; and Martinez, A. M. 2024. Latentpaint: Image inpainting in latent space with diffusion models. In *WACV*.

Dang, T.; Mascarich, F.; Khattak, S.; Papachristos, C.; and Alexis, K. 2019. Graph-based path planning for autonomous robotic exploration in subterranean environments. In *IEEE/RSJ IROS*.

Darabi, S.; Shechtman, E.; Barnes, C.; Goldman, D. B.; and Sen, P. 2012. Image melding: Combining inconsistent images using patch-based synthesis. *ACM TOG*.

Ding, G.; Zhao, C.; Wang, W.; Yang, Z.; Liu, Z.; Chen, H.; and Shen, C. 2024. FreeCustom: Tuning-Free Customized Image Generation for Multi-Concept Composition. In *CVPR*.

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

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 *WACV*.

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 *ICCV*.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*.

Greggman; and Engines, H. 2017. WebGL Aquarium — webglsamples.org. https://webglsamples.org/aquarium/aquarium.html. [Accessed 16-09-2024].

Guo, Q.; Gao, S.; Zhang, X.; Yin, Y.; and Zhang, C. 2017. Patch-based image inpainting via two-stage low rank approximation. *IEEE TVCG*.

Guo, X.; Yang, H.; and Huang, D. 2021. Image inpainting via conditional texture and structure dual generation. In *ICCV*.

He, Y.; Yang, S.; Chen, H.; Cun, X.; Xia, M.; Zhang, Y.; Wang, X.; He, R.; Chen, Q.; and Shan, Y. 2023. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *ICLR*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hu, Y.-T.; Chen, H.-S.; Hui, K.; Huang, J.-B.; and Schwing, A. G. 2019. SAIL-VOS: Semantic Amodal Instance Level Video Object Segmentation – A Synthetic Dataset and Baselines. In *CVPR*.

Jin, K. H.; and Ye, J. C. 2015. Annihilating filter-based low-rank Hankel matrix approach for image inpainting. *IEEE TIP*.

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 *ICCV*.

Labs, B. F. 2024. FLUX.1 [Schnell]. https://huggingface.co/black-forest-labs/FLUX.1-schnell. [Accessed 09-09-2024].

Lehar, S. 1999. Gestalt isomorphism and the quantification of spatial perception. *Gestalt theory*.

Li, J.; Wang, N.; Zhang, L.; Du, B.; and Tao, D. 2020. Recurrent feature reasoning for image inpainting. In *CVPR*.

Li, K.; and Malik, J. 2016. Amodal instance segmentation. In *ECCV*.

Li, W.; Lin, Z.; Zhou, K.; Qi, L.; Wang, Y.; and Jia, J. 2022. Mat: Mask-aware transformer for large hole image inpainting. In *CVPR*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, G.; Dundar, A.; Shih, K. J.; Wang, T.-C.; Reda, F. A.; Sapra, K.; Yu, Z.; Yang, X.; Tao, A.; and Catanzaro, B. 2022. Partial convolution for padding, inpainting, and image synthesis. *IEEE TPAMI*.

Liu, K.; Zhu, Z.; Li, C.; Liu, H.; Zeng, H.; and Hou, J. 2024. PrefPaint: Aligning Image Inpainting Diffusion Model with Human Preference. *arXiv preprint arXiv:2410.21966*.

Liu, Z.; Qiao, L.; Chu, X.; Ma, L.; and Jiang, T. 2025. Towards Efficient Foundation Model for Zero-shot Amodal Segmentation. In *CVPR*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*.

Martin, D.; Fowlkes, C.; Tal, D.; and Malik, J. 2001. A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *ICCV*.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *ICLR*.

Ozguroglu, E.; Liu, R.; Surś, D.; Chen, D.; Dave, A.; Tokmakov, P.; and Vondrick, C. 2024. pix2gestalt: Amodal Segmentation by Synthesizing Wholes. *CVPR*.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *ICCV*.

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 *CVPR*.

Qiu, H.; Xia, M.; Zhang, Y.; He, Y.; Wang, X.; Shan, Y.; and Liu, Z. 2024. Freenoise: Tuning-free longer video diffusion via noise rescheduling. In *ICLR*.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

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

Stability-AI. 2022. Stability-ai/stablediffusion: High-resolution image synthesis with Latent Diffusion Models. https://github.com/Stability-AI/stablediffusion. [Accessed 11-04-2024].

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*.

Tai, W.-E.; Shih, Y.-L.; Sun, C.; Wang, Y.-C. F.; and Chen, H.-T. 2025. Segment Anything, Even Occluded. In *CVPR*.

Tangemann, M.; Schneider, S.; Von Kügelgen, J.; Locatello, F.; Gehler, P.; Brox, T.; Kümmerer, M.; Bethge, M.; and Schölkopf, B. 2021. Unsupervised object learning via common fate. *arXiv preprint arXiv:2110.06562*.

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*.

Tschumperlé, D.; and Deriche, R. 2005. Vector-valued image regularization with PDEs: A common framework for different applications. *IEEE TPAMI*.

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*.

Xu, K.; Zhang, L.; and Shi, J. 2024. Amodal completion via progressive mixed context diffusion. In *CVPR*.

Yang, J.; Ren, Z.; Xu, M.; Chen, X.; Crandall, D. J.; Parikh, D.; and Batra, D. 2019. Embodied amodal recognition: Learning to move to perceive objects. In *ICCV*.

Yeh, R. A.; Chen, C.; Yian Lim, T.; Schwing, A. G.; Hasegawa-Johnson, M.; and Do, M. N. 2017. Semantic image inpainting with deep generative models. In *CVPR*.

Yildirim, A. B.; Pehlivan, H.; Bilecen, B. B.; and Dundar, A. 2023. Diverse inpainting and editing with gan inversion. In *ICCV*.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2019. Free-form image inpainting with gated convolution. In *ICCV*.

Yuille, A.; and Kersten, D. 2006. Vision as Bayesian inference: analysis by synthesis? *Trends in cognitive sciences*.

Zeng, Y.; Lin, Z.; Yang, J.; Zhang, J.; Shechtman, E.; and Lu, H. 2020. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *ECCV*.

Zeng, Y.; Patel, V. M.; Wang, H.; Huang, X.; Wang, T.-C.; Liu, M.-Y.; and Balaji, Y. 2024. JeDi: Joint-Image Diffusion Models for Finetuning-Free Personalized Text-to-Image Generation. In *CVPR*.

Zhan, G.; Zheng, C.; Xie, W.; and Zisserman, A. 2024. Amodal ground truth and completion in the wild. In *CVPR*.

Zhang, Z.; Chen, A.; Xie, L.; Yu, J.; and Gao, S. 2019. Learning semantics-aware distance map with semantics layering network for amodal instance segmentation. In *ACM MM*.

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

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