Amodal Completion via Progressive Mixed Context Diffusion

Katherine Xu[1] Lingzhi Zhang[2] Jianbo Shi[1]
[1]University of Pennsylvania, [2]Adobe Inc.

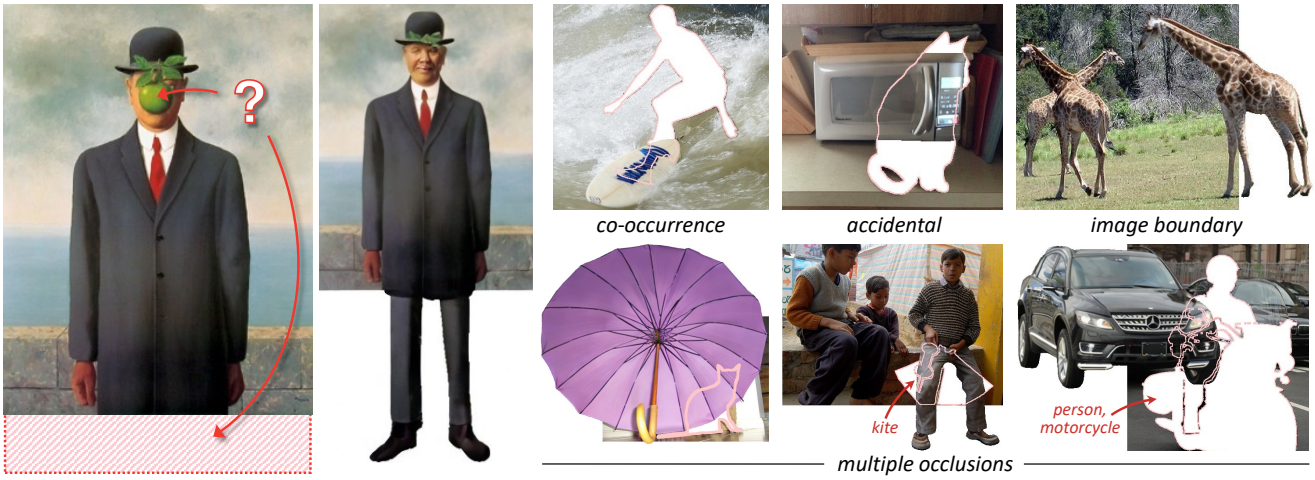

Figure 1. Our method can recover the hidden pixels of objects in diverse images. Occluders may be co-occurring (a person on a surfboard), accidental (a cat in front of a microwave), the image boundary (giraffe), or a combination of these scenarios.

Abstract

Our brain can effortlessly recognize objects even when partially hidden from view. Seeing the visible of the hidden is called amodal completion; however, this task remains a challenge for generative AI despite rapid progress. We propose to sidestep many of the difficulties of existing approaches, which typically involve a two-step process of predicting amodal masks and then generating pixels. Our method involves thinking outside the box, literally! We go outside the object bounding box to use its context to guide a pre-trained diffusion inpainting model, and then progressively grow the occluded object and trim the extra background. We overcome two technical challenges: 1) how to be free of unwanted co-occurrence bias, which tends to regenerate similar occluders, and 2) how to judge if an amodal completion has succeeded. Our amodal completion method exhibits improved photorealistic completion results compared to existing approaches in numerous successful completion cases. And the best part? It doesn't require any special training or fine-tuning of models.

1. Introduction

Have you ever wondered how objects regularly occlude one another, yet we can effortlessly recognize and imagine their unoccluded appearance? Our visual system performs this task of *amodal completion* using the continuity and symmetry of an object's shape [49] and everyday familiarity of the world [57]. Amodal completion, filling in hidden object parts, is a challenging AI task despite rapid advances in computer vision. This technology has many applications, such as in robotics, autonomous vehicles, and augmented reality.

A reasonable amodal completion approach [3, 60] consists of two stages: 1) completing a binary amodal mask; 2) synthesizing RGB pixel values within the mask. However, directly regressing the amodal mask is an ill-posed formulation due to the diversity of possible completions.

Computational issues aside, how can we create a dataset for amodal completion? Previous research attempts to construct datasets through random mask placement to simulate occlusion [24], computer graphics rendering techniques [17], or by asking humans to label amodal segmentation masks given the modal masks [1, 39, 67]. However, a domain gap persists between synthetic and natural images, and labeled natural images are expensive to obtain.

But what if we can sidestep all these difficulties? This involves thinking outside the box, literally: 1) extend from the bounding box of an occluded object to include sufficient image context, 2) remove the occluders, 3) use a pre-trained diffusion model to grow the object, and 4) trim off the extra background. Our amodal completion pipeline avoids pre-

Project page and code: https://k8xu.github.io/amodal/

dicting the amodal segmentation mask as an intermediate step. Furthermore, we can recover occluded pixels within occluder objects and *beyond the image boundary* (Figure 1), generate *diverse* versions of completed pixels, and require *no training or fine-tuning* of the diffusion model.

Straightforward usage of an off-the-shelf diffusion model for image inpainting succeeds only sometimes. Failure cases often generate other objects within the occluder masks: original occluder look-alikes or unintended things that co-occur with the object of interest. Imagine removing a hand holding a cup; diffusion inpainting often adds a different hand simply because we don't see a floating cup in real life.

How can we add control to discourage the pre-trained diffusion model from re-generating co-occurrence? Two tasks exist to solve: momentarily breaking free of context, and knowing when amodal completion has succeeded.

First, to break free of the contextual bias that causes co-occurrence, we propose *mixed context diffusion sampling* to temporarily replace the image context with a natural clean background, akin to product photography. We intercept the diffusion process halfway, extracting a pseudo-complete object in a still-noisy image using unsupervised clustering of decoder features. Then, we use the pseudo-complete object as a reference target to the reverse diffusion process while gradually reintroducing the original image background.

Second, to infer whether the amodal completion succeeds, we introduce counterfactual reasoning: use the generated object and outpaint its background. If the object is complete, then any outpainting should not increase its size. Thus, we can judge whether amodal completion is successful by comparing the object segmentation before and after outpainting.

With these two tools, we progressively run a pre-trained text-based diffusion inpainting model [42] until the occluded object is complete. Our method requires no extra datasets, retraining, or adaptation. It is entirely based on pre-trained diffusion models [42], complemented by off-the-shelf grounded segmentation [18, 29] and depth models [21] as auxiliary modules. In summary:

1. We introduce a *progressive occlusion-aware amodal completion pipeline* that effectively recovers hidden pixels within occluder masks and beyond the image boundary.
2. As a pioneering exploration, we identify the challenge of a diffusion inpainting model generating unwanted co-occurring objects, instead of completing objects. We propose *mixed context diffusion sampling*, which modifies the image context to overcome difficult co-occurrence.
3. We create a training-free *counterfactual completion curation system* to decide if a generated object is complete.

2. Related Work

Amodal completion. Amodal appearance completion aims to fill in the hidden regions of occluded objects. Current methods often rely on a two-step approach of first predicting the amodal mask and then generating the object appearance given the amodal mask. These approaches have been applied on toy datasets [4, 10, 13] and specific object categories such as vehicles [27, 54, 65], humans [66], and food [37]. Additional methods perform amodal completion for common object categories, mainly on synthetic indoor scenes [7, 9, 65]. However, there is a domain gap between synthetic and natural images, and so we are interested in generating the amodal completion of common objects in natural scenarios. Zhan et al. [60] and Bowen et al. [3] perform amodal completion in natural images, but these GAN-based methods tend to lack high image fidelity. In contrast, our approach leverages the good image prior of pre-trained diffusion models to photorealistically complete objects in natural images. We also note the presence of concurrent works [16, 36, 58].

An alternative means of tackling the amodal completion task is directly inpainting occluder regions, such as by using large mask image inpainting [47] or training diffusion inpainting models to remove objects [56]. There is also a line of research in image outpainting [6, 23, 25, 51–53]. However, since these approaches are not meant for amodal completion, they often produce realistic images but fail to complete the appearance of desired objects under significant occlusion. In contrast, our method can realistically generate the amodal completion of occluded objects by progressively inpainting occluder regions and disentangling co-occurrence bias.

Diffusion models. Inspired by non-equilibrium thermodynamics [45], diffusion models achieve remarkable results for text-to-image and image inpainting tasks [34, 41, 42, 44], often outperforming GANs in generating photorealistic and diverse images [8, 15, 33]. However, these approaches provide limited control of image generation outside the text prompt. Several works provide additional guidance to diffusion models for controllable image generation using CLIP [34, 41], cross-attention and self-attention [5, 11, 14, 38], stroke paintings [31], exemplar images [55], and extra conditioning on segmentation maps, edge maps, bounding boxes, and keypoints in ControlNet [62]. However, employing these techniques for amodal completion typically needs an amodal mask or 'complete' edge map as guidance, which is not available. Moreover, they often require resource-intensive retraining of the diffusion models. In contrast, our method does not assume any initial guidance and is entirely training-free.

3. Method

3.1. Preliminaries

Diffusion models [15] learn a data distribution $p(I)$ using a sequence of denoising. In the forward process, the model adds noise to an image I in N time steps, resulting in the sample having approximately Gaussian noise. In the reverse process, the model learns to denoise the sample in N steps. At each step $t = [1, N]$, a learned neural network predicts the noise $\epsilon_\theta(I^t, t)$ given the noisy image I^t.

Unlike diffusion models that use the image pixel space, la-

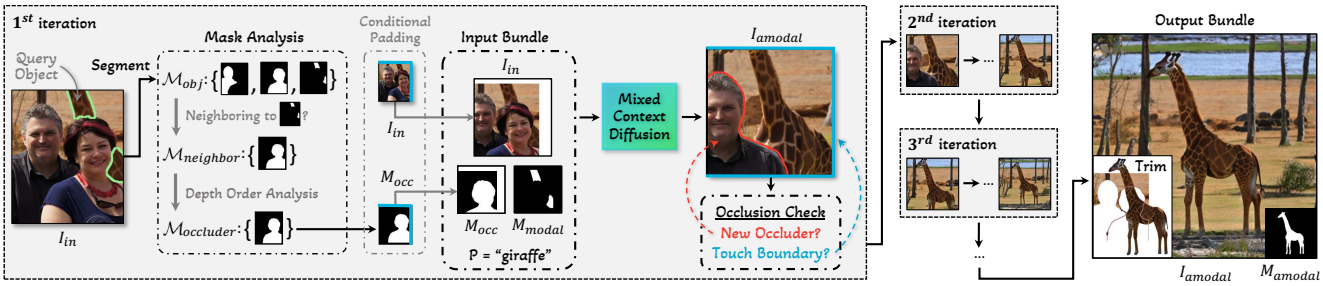

Figure 2. Our **Progressive Occlusion-aware Completion** pipeline. **First iteration:** We perform instance segmentation [18, 29] and analyze the object masks to determine occluders [21]. If the query object touches the image boundary, then we pad the image and mask to enable object completion beyond the boundary in those directions. Using this input bundle, we run our Mixed Context Diffusion Sampling to obtain a new amodal completion image. *The details of Mixed Context Diffusion Sampling are in Figure 4.* Next, we check whether the generated object has a new occluder or touches the image boundary. In this example, the man from the original image appears as a new occluder that was previously undetected. **Additional iterations:** If the query object remains occluded, then we run additional iterations of our pipeline. **Output:** We return the final amodal completion image and amodal mask, and we can trim extra background to overlay on the original image.

tent diffusion models (LDMs) [42] operate in the latent space of pre-trained autoencoders. For an image $I \in \mathbb{R}^{H\times W\times 3}$, the encoder E encodes I into the latent representation $\mathbf{z} = E(I)$, and the decoder D reconstructs I from $\mathbf{z}$ using $\hat{I} = D(\mathbf{z})$. In the diffusion process, the autoencoder can be viewed as a time-conditional UNet [43], $\epsilon_\theta(\mathbf{z}_t, t)$, for a given time t and latent $\mathbf{z}_t$. To incorporate any input condition y such as an inpainting mask, LDMs add cross-attention layers [50] to the denoising UNet so that y maps to the intermediate layers of the UNet [42]. In this work, we use the publicly released Stable Diffusion v2 inpainting model checkpoint [42]. For simplicity, *our notation uses the image pixel space hereafter*.

3.2. Problem Setup

The task of amodal completion entails identifying both visible and hidden aspects of objects, inside and outside an image's boundary. Given an arbitrary image I_{in} in $\mathbb{R}^{H\times W\times 3}$ and an object of interest ('query object') with its modal mask M_{modal} in $\mathbb{R}^{H\times W}$, our objective is to predict the amodal completion image I_{amodal} in $\mathbb{R}^{H'\times W'\times 3}$ and the corresponding amodal mask M_{amodal} in $\mathbb{R}^{H'\times W'}$. We use H' and W' to indicate that the final amodal completion image may differ in size from the original image due to potential extensions beyond the image boundary. Furthermore, we denote the diffusion inpainting/outpainting process as $F_{s\to e}$, where s is the starting timestep and e is the ending timestep. The text prompt is represented as P, which is the semantic category of the query object for our proposed solutions. M_{in} and I_{out} signify the generic input mask and the generated output, respectively. This diffusion process can be expressed as:

$$I_{out} = F_{s\to e}(I_{in}, M_{in}, P) \tag{1}$$

For amodal completion to be successful, it must fulfill three criteria. 1) The process should exclusively remove occluders without altering the image background, thereby avoiding overextension of the object. 2) It must ensure a complete representation of all object parts to avoid incompletion. 3) The completion must be contextually consistent, avoiding any physically implausible object configurations. We evaluate the first two criteria by developing a dataset of unoccluded objects from natural images, and then artificially generating the pseudo-occluded versions. Contextual consistency is harder to quantify, so we perform a user study.

3.3. Naive Outpainting Approach

A simple approach to amodal completion may assume that all pixels outside the query object's modal mask are occlusions, which are then subject to outpainting. This 'Naive Outpainting' approach can be mathematically expressed as:

$$I_{amodal} = F_{0\to N}(I_{in}, 1 - M_{modal}, P) \tag{2}$$

where N is the total number of timesteps for the diffusion process, which is set to 50 in the DDIM scheduler [46]. We treat all masks as binary, so $1 - M_{modal}$ signifies everything exterior to the query object.

Naive Outpainting often overextends the query object due to the lack of contextual constraints, compromising the integrity of its identity and violating the objective of amodal completion. For example, this approach produces an unwanted change to the motorcycle's orientation in Figure 3.

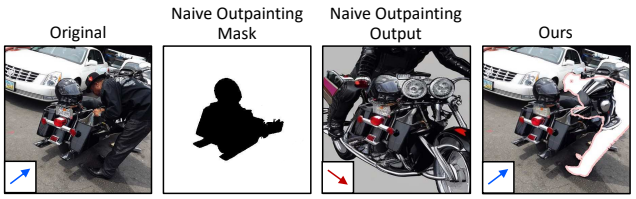

Figure 3. Naively using a diffusion model to outpaint the query object may overextend the object and change its identity, such as the motorcycle changing orientation (indicated by arrow). In contrast, our method preserves the object's identity.

3.4. Progressive Occlusion-aware Completion

Our method is based on two key insights: 1) inpainting only where necessary by identifying occluders prevents overextension, and 2) iteratively performing this inpainting step avoids incompletion. Thus, we propose a 'Progressive Occlusion-aware Completion' pipeline, as shown in Figure 2. Each

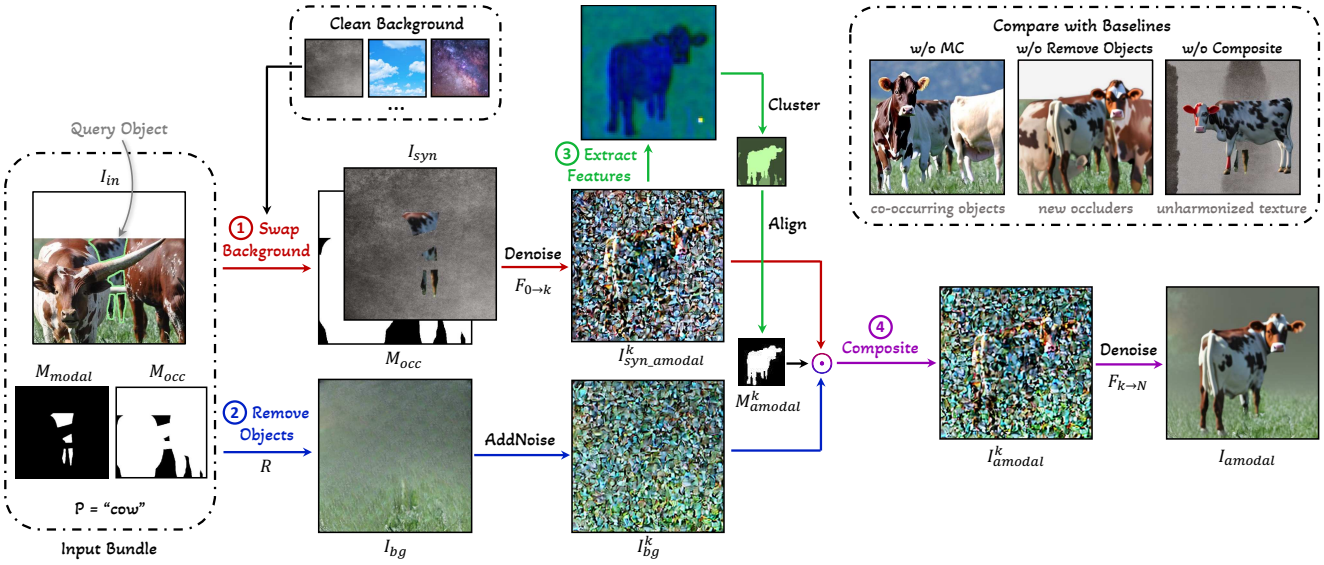

Figure 4. Our **Mixed Context (MC) Diffusion Sampling**. 1) **Swap background** (*red*): We replace the background of I_{in} using M_{occ} to create I_{syn}, followed by diffusion inpainting to the k^{th} timestep, resulting in $I^k_{syn\_amodal}$. 2) **Create object-removed background image** (*blue*): We remove query objects and occluders from I_{in} using a removal inpainter [47], and then add noise up to the k^{th} timestep, producing I^k_{bg}. 3) **Segment object in noisy image** (*green*): We extract diffusion features from $I^k_{syn\_amodal}$, cluster them, and select the query object's amodal mask M^k_{amodal} at the k^{th} timestep by aligning with M_{modal}. 4) **Composite** (*purple*): We use M^k_{amodal} to place the query object from $I^k_{syn\_amodal}$ onto the object-removed background image I^k_{bg}. The final image, I_{amodal}, is obtained by completing the remaining $N-k$ diffusion steps, where N is the total number of steps. **Top right:** We show various failure cases if we remove parts of this MC method.

iteration of our pipeline has several steps that are described below, and we perform more iterations if occluders remain. This approach significantly reduces object overextension, as evidenced by the results on the right side of Figure 6.

Mask analysis. Given the input image I_{in}, the first step in each iteration is to identify all object masks by applying a grounded segmentation model [18, 29]. This set of masks is denoted as $\mathcal{M}_{obj} = \{M_1, M_2, \ldots, M_n\}$, with each M_i representing a distinct object mask. To focus on the objects neighboring the query object mask M_{modal}, we filter $\mathcal{M}_{obj}$ to yield $\mathcal{M}_{neighbor} = \{M_1, M_2, \ldots, M_j\}$.

We also perform a depth ordering analysis [21] between M_{modal} and the masks in $\mathcal{M}_{neighbor}$, and we consider any mask in $\mathcal{M}_{neighbor}$ that is closer to the camera than M_{modal} as an occluder. Next, this set of occluder masks $\mathcal{M}_{occluder}$ is aggregated into a single binary occlusion mask, mathematically expressed as $M_{occ} = \sum_{M_i \in \mathcal{M}_{occluder}} M_i$. This unified mask M_{occ} captures occluders within the image boundary and serves as the input mask for the diffusion process.

Conditional padding. Our approach completes objects that may extend beyond the image boundary by including the boundary as an occluder. If the query object mask M_{modal} touches the boundary, then we apply padding to the image I_{in} and the input mask M_{occ} in the corresponding directions.

Diffusion process and occlusion check. After mask analysis and conditional padding, we zoom into the query object by cropping I_{in}, M_{occ}, and M_{modal} around its bounding box. This can improve the image generation quality by the diffusion inpainting process, as described in [63].

The input bundle to the diffusion process contains the new image I_{in}, occluder mask M_{occ}, query object's modal mask M_{modal}, and semantic category for P. We run our Mixed Context Diffusion method (Section 3.5) and generate a new amodal completion image I_{amodal} using the equation:

$$I_{amodal} = F_{0\to N}(I_{in}, M_{occ}, P) \tag{3}$$

At the end of each pipeline iteration, we check if the object is still occluded by other objects or the image boundary.

Additional iterations. If occlusions remain, then we run another iteration of our pipeline using the previous iteration's amodal completion image I_{amodal} as the new input I_{in}, and the previous iteration's amodal mask M_{amodal} as the new modal mask M_{modal}. Our pipeline continues until the query object is no longer occluded. Lastly, we return an output bundle with the final amodal completion image and amodal mask. To visualize the completed object in the original context, we can trim the extra background from I_{amodal} and overlay the object on the original image.

3.5. Mixed Context Diffusion Sampling

Our Progressive Occlusion-aware Completion pipeline involves inpainting occluder regions, but directly using a pre-trained diffusion inpainting model may generate co-occurring objects as new occluders due to contextual bias. This bias extends to subtle details like shadows, which can prompt the model to produce contextually compatible occluders in the edited region, as discussed in [59].

To address this, we temporarily break the co-occurrence link between the query object and original image context

during the diffusion process. We achieve this through our 'Mixed Context Diffusion Sampling' (MC), presented in Figure 4. This approach is versatile and can be adapted to any text-to-image diffusion model in pixel or latent space.

After receiving the input bundle described in Section 3.4, our approach bifurcates into two parallel paths. The first path aims to complete the query object by reducing contextual bias, while the second path frees the original image of occluders. Then, we composite the noisy images from both paths into a single noisy image by creating an intermediate query object mask. We explain each step below.

Swap background. We replace the area of I_{in} outside the query object's modal mask M_{modal} with a clean background, reminiscent of the gray backdrops typically used in product photography. This creates a synthetically composited image, denoted as I_{syn}. Next, in the Denoise step, we apply diffusion inpainting using I_{syn} and M_{occ} up to the k^{th} diffusion timestep. This can be mathematically expressed as:

$$I^k_{syn\_amodal} = F_{0\to k}(I_{syn}, M_{occ}, P) \tag{4}$$

Create object-removed background image. We produce a clean background image of the original context, devoid of both query and occluder objects. To achieve this, we first remove them from I_{in} using the combined area of $M_{modal} + M_{occ}$ via a removal inpainter R, as described in [47]. The output from R is then subjected to noise addition using AddNoise$(\cdot, k)$ up to the k^{th} timestep. This results in I^k_{bg}, which is the noise-infused clean background image after k^{th} timesteps. This can be mathematically expressed as:

$$I^k_{bg} = \text{AddNoise}(R(I_{in}, M_{modal} + M_{occ}), k) \tag{5}$$

Segment query object in noisy image. After deriving $I^k_{syn\_amodal}$ and I^k_{bg}, each characterized by the k^{th} noise level, we aim to insert the query object from $I^k_{syn\_amodal}$ into the object-removed background image. Central to this step is the determination of an appropriate intermediate query object mask from the noisy image $I^k_{syn\_amodal}$.

Our insight is that segmenting the query object from the noisy image $I^k_{syn\_amodal}$ is difficult, but we can use the latent information from the UNet decoder to find clusters [30] for query object mask proposals. Our clustering approach is similar to [2, 22]. We experimentally determined the best l^{th} decoder layer and k^{th} timestep to extract features. Each cluster is associated with different segments of the image. We compute pixel overlap of each cluster with the modal mask M_{modal} to select the segment that best aligns with the query object in the noisy image $I^k_{syn\_amodal}$. This segment is the amodal object mask M^k_{amodal} at the k^{th} timestep.

Composite. We use M^k_{amodal} to composite the query object back onto the object-removed background image I^k_{bg}, instead of the original image, to ensure that the completed query object is contextually consistent. We create this composited image I^k_{amodal} as follows:

$$I^k_{amodal} = I^k_{syn\_amodal} \odot M^k_{amodal} + I^k_{bg} \odot (1 - M^k_{amodal})$$

Finally, we continue the diffusion process for $N - k$ steps using the composited image I^k_{amodal} and the occluder mask M_{occ}. We denote this remaining diffusion process as $F_{k\to N}$ and obtain the final image I_{amodal}. This is expressed as:

$$I_{amodal} = F_{k\to N}(I^k_{amodal}, M_{occ}, P) \tag{6}$$

3.6. Counterfactual Completion Curation System

After generating a set of amodal completion images, how can we decide if the objects are successfully completed? Inspired by [35], we propose a 'Counterfactual Completion Curation System' that can reduce the burden of human labeling by filtering unsuccessful completions without model training. The intuition behind our system is that outpainting incomplete objects is more likely to generate more pixels belonging to missing object parts than outpainting complete objects. Our initial curation system relies on a training-free rule to classify generated objects as complete or incomplete.

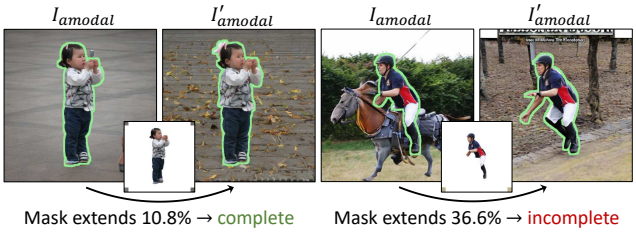

Figure 5. Our counterfactual completion curation system uses a training-free rule to determine complete and incomplete objects. We outpaint the object everywhere except the image corners, and then compare amodal masks from I_{amodal} and I'_{amodal}. We experimentally determined a mask extension threshold of 20%.

Figure 5 shows complete and incomplete objects determined by our rule, which has a generation step and a decision step. In the generation step, we outpaint the object in I_{amodal} using an input mask M_{in} consisting of everywhere except the amodal mask M_{amodal} and the image corners, which guides the diffusion process towards a reasonable background. Then, we trim the background in the new amodal completion image I'_{amodal} to extract the query object and obtain a new amodal mask M'_{amodal}. In the decision step, we classify objects using thresholds on two parameters: 1) the object's proximity to the image boundary, and 2) the extension of the amodal mask area. To set the two thresholds, we use a small validation set of 100 images.

4. Experiments

Our Progressive Occlusion-aware Completion pipeline and Mixed Context Diffusion Sampling can successfully fill in hidden object pixels in a variety of object categories and

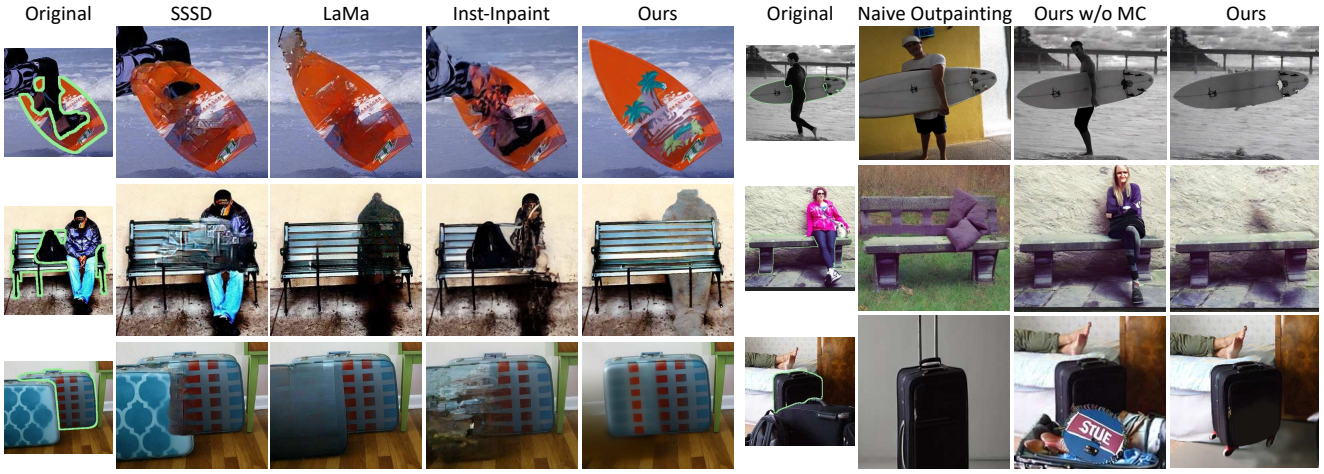

Figure 6. **Left:** Comparison of our method with prior works on natural images. **Right:** Comparison of our method, our method without Mixed Context Diffusion Sampling (MC), and Naive Outpainting. Our method extends objects only where necessary unlike Naive Outpainting. Additionally, our approach avoids generating co-occurring objects, unlike ours without MC and Naive Outpainting.

occlusion cases in natural images. Notably, it can complete objects inside and outside the image boundary, overcome difficult co-occurrence bias, and handle high occlusion rates using the pre-trained diffusion model's good image prior.

4.1. Comparisons with Previous Methods

Prior Works. We compare with three prior works: a GAN-based amodal completion method *Self-Supervised Scene Deocclusion* (SSSD) [60], a GAN-based inpainting method *Large Mask Inpainting* (LaMa) [47], and a diffusion-based object removal method *Inst-Inpaint* [56]. For fair comparison, we focus on completing objects within the image boundary because they do not extend the boundary. We use LaMa to fill in occluders and Inst-Inpaint to remove occluders.

Datasets. One main challenge of evaluating our amodal completion task is that there are no natural image datasets with ground truth amodal appearance completions for common object categories. In addition, existing amodal datasets with ground truth amodal masks do not consider the diversity of possible object completions. To bypass these limitations, we create a dataset of 3,000 pseudo-occluded common objects and their completed counterparts using natural images from COCO [26] and Open Images [19, 20]. As shown in Figure 7, we simulate occlusion by overlaying a complete object on the image of another complete object. Our dataset contains diverse, challenging scenarios for amodal completion, covering at least 55 object categories with significant occlusion rates: 1,500 easy cases with 20-50% occlusion and 1,500 hard cases with 50-80% occlusion. In Figure 6, we present qualitative results on natural images.

Metrics. We assess the quality of the generated amodal completion images with the ground truth complete object images at three image similarity levels. We use CLIP [40] for high-level, DreamSim [12] for mid-level, and LPIPS [64] for low-level. For CLIP, we compute the cosine similarity between image embeddings of the generated amodal completion and text embeddings of the query object category. For DreamSim and LPIPS, we calculate the perceptual distance between generated and ground truth complete object image. We segment and place the completed objects on a black background to focus on the query object appearance. Furthermore, we conduct a user preference study to evaluate how well the generated object fits into its original context. The right side of Figure 7 provides an example of the various aspects of the amodal completion image to evaluate.

Figure 7. **Left:** We place complete objects on each other to create pseudo-occluded objects. Here, the bus is 66.2% occluded by the donut. **Right:** We can evaluate whether A' (formerly occluded) successfully completes the query object A by using pseudo-occluded objects and off-the-shelf metrics. But, these metrics do not assess whether $A' \cup A$ fits into the background B. To this end, we conduct a user preference study to evaluate the generated objects in context.

Quantitative Results. In Table 1, we report the mean image similarity scores across all pseudo-occluded object images. Our method generally performs better than prior works for both easy cases and hard cases. Interestingly, LaMa [47] obtains similar scores to our method even though it is not intended for amodal completion. We suspect that LaMa seems to perform well because it often generates similarly colored pixels as the query object within the inpainted region, and a visual assessment of LaMa's output verifies that it creates blurry object appearances and boundaries. For this reason, we conduct a user preference study and present qualitative results to further measure the amodal completion quality between our method and prior works.

User Preference Study. We conduct a user preference

Method	Easy Cases			Hard Cases			User Preference
	CLIP ↑	DreamSim ↓	LPIPS ↓	CLIP ↑	DreamSim ↓	LPIPS ↓	
SSSD [60]	0.280 / 0.263	0.186 / 0.216	0.096 / 0.142	0.267 / 0.263	0.315 / 0.334	0.166 / 0.225	1.8%
LaMa [47]	0.288 / 0.265	0.098 / 0.124	**0.054** / 0.091	0.279 / **0.268**	0.236 / 0.292	0.130 / 0.205	7.3%
Inst-Inpaint [56]	0.264 / 0.257	0.325 / 0.304	0.185 / 0.195	0.252 / 0.254	0.451 / 0.446	0.263 / 0.283	0.0%
Ours	**0.290 / 0.266**	**0.096 / 0.106**	**0.054 / 0.078**	**0.290** / 0.267	**0.184 / 0.185**	**0.110 / 0.141**	**90.9%**

Table 1. Our method overall performs better than prior works in terms of CLIP (high-level), DreamSim (mid-level), and LPIPS (low-level) image similarity. We use 2,500 objects from COCO and 500 objects from Open Images, and *scores are formatted as COCO / Open Images*. We consider easy cases where the object has 20-50% occlusion and hard cases with 50-80% occlusion. Additionally, we observe that users highly prefer the generated amodal completions using our method across 55 easy cases and 55 hard cases from COCO. The user preference percentages are coincidentally the same for easy and hard cases.

study to more accurately assess the perceptual image quality of the generated amodal completions. We randomly select 55 easy cases and 55 hard cases of pseudo-occlusion images and solicit feedback from at least three Amazon Mechanical Turk (MTurk) workers. We show the pseudo-occluded object image and the generated object images from each method side by side, and we ask each worker to vote on the generated object that looks most complete and realistic. As shown in Table 1, the user preferences demonstrate that our method significantly outperforms prior works in the visual quality of amodal completion images.

Qualitative Results. Figure 6 visually compares our method and prior works on occluded objects in natural images. We notice that SSSD [60] often generates visual artifacts and incomplete objects, LaMa [47] produces ill-defined object boundaries and unrealistic appearances, and Inst-Inpaint [56] can remove objects but struggles to complete them. Our method can complete highly occluded objects with realistic and contextually consistent appearances.

Implementation. We demonstrate our method using the publicly available Stable Diffusion v2 inpainting model [42]. All experiments use a 24GB Nvidia Titan RTX GPU, and our method does not involve any training or fine-tuning.

4.2. Ablation Studies

We ablate our amodal completion method to demonstrate the effectiveness of our Progressive Occlusion-aware Completion pipeline and Mixed Context Diffusion Sampling. We randomly select 100 occluded objects from natural images and generate their completed versions using our method and Naive Outpainting. We consider 50 hard cases where the occluder is the top co-occurring semantic category for the query object, and 50 easy cases where the occluder is not.

We conduct a user study to find the number of successful amodal completions for each method by soliciting feedback from at least three MTurk workers. In addition, we perform a user preference study and ask at least six MTurk workers to vote on the method that generates the most complete and realistic objects. In Table 2, our method outperforms Naive Outpainting in successful completions and user preference, even without Mixed Context Diffusion Sampling. On hard cases, we observe that using Mixed Context Diffusion Sampling significantly aids successful completions by +18%.

Method	Easy Cases		Hard Cases	
	Successes	User Preference	Successes	User Preference
Naive Outpainting	66%	18%	40%	18%
Ours w/o MC	**90%**	36%	72%	28%
Ours	88%	**48%**	**90%**	**54%**

Table 2. Ablation study of amodal completion successes and user preference. We use 50 easy cases and 50 hard cases, where the occluder is the top co-occurring object category.

Figure 6 visually compares each method on natural images with challenging co-occurrence or highly occluded objects. Our method prevents co-occurring objects and inpaints only where necessary, compared to ours without Mixed Context Diffusion Sampling and Naive Outpainting.

4.2.1 Mixed Context Diffusion Sampling

We experiment with the clean image to swap background, the UNet layer to cluster features and segment the query object in the noisy image, and the timestep to composite the query object on the object-removed background image.

Clean image to swap background. We test the effect of five clean backgrounds to swap with the original image background on a small set of 55 images. On the left of Figure 8, using a gray background led to a +20% increase in successful amodal completions, while using a forest or sky background led to a noticeable drop in completion performance (-16% and -9%) compared to using the original context.

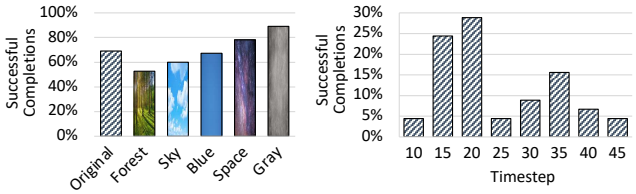

Figure 8. **Left:** Using a gray background improves completion by +20% compared to the original background. **Right:** Swapping contexts at DDIM timestep 20 out of 50 leads to more successful completions on *difficult co-occurrence cases*.

UNet layer to segment object in noisy image. To cluster features and create the intermediate amodal mask M^k_{amodal}, we examine the effect of using different UNet layers in Figure 9. We observe that features from the third UNet decoder layer best capture object geometry and low-level visual features for Mixed Context Diffusion Sampling, as described in [32, 48]. Early encoder layers sometimes cluster the inpaint-

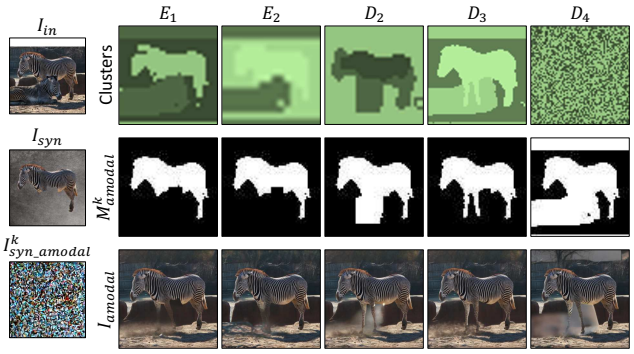

Figure 9. We analyze how features from different UNet encoder and decoder layers affect the clusters, intermediate amodal mask M^k_{amodal}, and final amodal completion image I_{amodal}. Here, we show a subset of the UNet layers, and E_l and D_l refer to the l-th layer of the encoder and decoder, where $l \in [1, 2, 3, 4]$. The leftmost column shows the input image I_{in} for the current iteration of our pipeline, the synthetically overlaid input image I_{syn}, and the intermediate inpainted object $I^k_{syn\_amodal}$. We discover that features from D_3 generally produce good clusters to create a well-defined M^k_{amodal}. Features from other layers often cannot fully capture thin structures, such as the zebra's legs in this example.

ing mask region, and the last decoder layer is often noisy. The second decoder layer sometimes captures the general object shape but is less fine-grained. Despite poor clusters and intermediate amodal mask, the final amodal completion of the query object is not always negatively impacted.

Timestep to composite. On the right of Figure 8, we analyze the timestep to swap image backgrounds on 45 images where our method without Mixed Context Diffusion Sampling generates a co-occurring occluder or unwanted visual artifacts. We record the earliest timestep that returns a good intermediate amodal mask M^k_{amodal} and final amodal completion image I_{amodal}. We observe the peak at timestep 20, indicating that object shape may appear relatively early on during the denoising process. Nonetheless, the ideal timestep for compositing the query object onto the object-removed background image often depends on the occlusion level.

4.3. Counterfactual Completion Curation System

Our curation system involves a training-free rule that classifies generated objects as complete or incomplete. On the validation set of 100 images, our rule reaches a mean accuracy of 0.73, precision of 0.73, and recall of 0.71 from 3 different trials. We evaluate our rule by measuring accuracy, precision, and recall on a test set of 50 complete objects and 50 incomplete objects. As shown in Table 3, our rule can achieve an accuracy of 0.70, despite not being specially trained on any datasets. There is room for improvement, which can be mitigated by training models on a curated dataset of complete and incomplete objects.

Furthermore, we compare the performance of our rule with that of humans. We ask three humans to independently classify the 100 test images, and we compute the human consensus using a simple majority vote for each object. Human consensus results in an accuracy of 83%. This indicates the subjective nature of this binary classification task and further shows the need for an accurate and reliable curation system.

	Accuracy	Precision	Recall
Human Consensus	0.83	0.75	0.99
Our Rule	0.70	0.68	0.68
Random Chance	0.50	0.50	0.50

Table 3. Comparison of our counterfactual rule with humans on classifying 100 test images as complete or incomplete. We compare the mean scores of our rule from three different trials with the human consensus as a simple majority vote from three humans.

5. Discussion

We introduced a new approach for amodal completion using diffusion inpainting. Our method progressively inpaints obscured regions by analyzing occlusion for a query object. Unlike conventional two-step approaches that predict the amodal mask and then complete the amodal appearance, ours is the first to predict the appearance directly. We deploy mixed context diffusion sampling to reduce the inpainting of unintended co-occurring objects. We also establish a counterfactual-based curation system for measuring object completeness. With our method, we can build dense correspondence [61] and 3D novel view synthesis [28] on highly occluded visual objects, as illustrated in Figure 10.

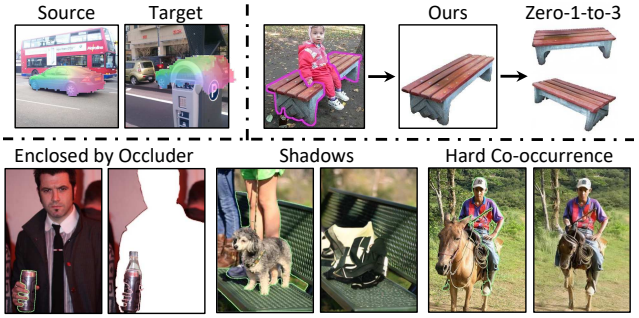

Figure 10. **Top:** Two applications of our amodal completion method are dense correspondence [61] and novel view synthesis [28]. **Bottom:** One failure case occurs when the object is enclosed by the occluder, leading to overextension. Additionally, our method sometimes struggles to complete objects due to shadows or the object's pose, such as the person riding a horse.

Limitations. Our method, while effective, encounters limitations with everyday occlusions, as shown in Figure 10. Challenges arise when a small query object is obscured by a larger occluder, potentially leading to its overextension. Additionally, subtle shadows on the query object can inadvertently introduce compatible occluders. Moreover, certain human poses strongly suggest interaction with other objects, occasionally resulting in the generation of unintended occluders. Despite these challenges, we believe our method establishes a solid new benchmark in amodal completion. Addressing these complex scenarios remains an exciting avenue for future work.

References

[1] Jiayang Ao, Qiuhong Ke, and Krista A. Ehinger. Amodal intra-class instance segmentation: New dataset and benchmark, 2023. 1

[2] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models, 2022. 5

[3] Richard Strong Bowen, Huiwen Chang, Charles Herrmann, Piotr Teterwak, Ce Liu, and Ramin Zabih. Oconet: Image extrapolation by object completion. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2307–2317, 2021. 1, 2

[4] Christopher P. Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation, 2019. 2

[5] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance, 2023. 2

[6] Yen-Chi Cheng, Chieh Hubert Lin, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, and Ming-Hsuan Yang. Inout: Diverse image outpainting via gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11431–11440, 2022. 2

[7] Helisa Dhamo, Nassir Navab, and Federico Tombari. Object-driven multi-layer scene decomposition from a single image, 2019. 2

[8] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. 2

[9] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible, 2018. 2

[10] Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. Genesis: Generative scene inference and sampling with object-centric latent representations, 2020. 2

[11] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis, 2023. 2

[12] Stephanie Fu*, Netanel Tamir*, Shobhita Sundaram*, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv:2306.09344*, 2023. 6

[13] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Chris Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference, 2020. 2

[14] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 2

[16] Cheng-Yen Hsieh, Tarasha Khurana, Achal Dave, and Deva Ramanan. Tracking any object amodally, 2024. 2

[17] Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, and Alexander G Schwing. Sail-vos: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3105–3115, 2019. 1

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 2, 3, 4

[19] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017. 6

[20] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 6

[21] Hyunmin Lee and Jaesik Park. Instance-wise occlusion and depth orders in natural scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21210–21221, 2022. 2, 3, 4

[22] Daiqing Li, Huan Ling, Amlan Kar, David Acuna, Seung Wook Kim, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Dreamteacher: Pretraining image backbones with deep generative models, 2023. 5

[23] Jiacheng Li, Chang Chen, and Zhiwei Xiong. Contextual outpainting with object-level contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11451–11460, 2022. 2

[24] Ke Li and Jitendra Malik. Amodal instance segmentation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 677–693. Springer, 2016. 1

[25] Yijun Li, Lu Jiang, and Ming-Hsuan Yang. Controllable and progressive image extrapolation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2140–2149, 2021. 2

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 6

[27] Huan Ling, David Acuna, Karsten Kreis, Seung Wook Kim, and Sanja Fidler. Variational amodal object completion. In *Advances in Neural Information Processing Systems*, pages 16246–16257. Curran Associates, Inc., 2020. 2

[28] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 8

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with

grounded pre-training for open-set object detection, 2023. 2, 3, 4

[30] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. 5

[31] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2022. 2

[32] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models, 2023. 7

[33] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. 2

[34] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022. 2

[35] Deniz Oktay, Carl Vondrick, and Antonio Torralba. Counterfactual image networks. 2018. 5

[36] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes, 2024. 2

[37] Dim P. Papadopoulos, Youssef Tamaazousti, Ferda Ofli, Ingmar Weber, and Antonio Torralba. How to make a pizza: Learning a compositional layer-based gan model, 2019. 2

[38] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models, 2023. 2

[39] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2019. 1

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 6

[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 2

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 7

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 3

[44] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 2

[45] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. 2

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[47] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions, 2021. 2, 4, 5, 6, 7

[48] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion, 2023. 7

[49] Robert van Lier. Investigating global effects in visual occlusion: from a partly occluded square to the back of a tree-trunk. *Acta Psychologica*, 102:203–220, 1999. 1

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[51] Xi Wang, Weixi Cheng, and Wenliang Jia. Structure-guided image outpainting, 2022. 2

[52] Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[53] Qingguo Xiao, Guangyao Li, and Qiaochuan Chen. Image outpainting: Hallucinating beyond the image. *IEEE Access*, 8:173576–173583, 2020. 2

[54] Xiaosheng Yan, Yuanlong Yu, Feigege Wang, Wenxi Liu, Shengfeng He, and Jia Pan. Visualizing the invisible: Occluded vehicle segmentation and recovery, 2019. 2

[55] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models, 2022. 2

[56] Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegul Dundar. Inst-inpaint: Instructing to remove objects with diffusion models, 2023. 2, 6, 7

[57] Xuyan Yun, Simon Hazenberg, and Rob Lier. Temporal properties of amodal completion: Influences of knowledge. *Vision research*, 145, 2018. 1

[58] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. Amodal ground truth and completion in the wild, 2023. 2

[59] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. What does stable diffusion know about the 3d scene? *arXiv preprint arXiv:2310.06836*, 2023. 4

[60] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion, 2020. 1, 2, 6, 7

[61] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. 2023. 8

[62] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2

[63] Lingzhi Zhang, Zhengjie Xu, Connelly Barnes, Yuqian Zhou, Qing Liu, He Zhang, Sohrab Amirghodsi, Zhe Lin, Eli Shecht-

man, and Jianbo Shi. Perceptual artifacts localization for image synthesis tasks, 2023. 4

[64] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. 6

[65] Chuanxia Zheng, Duy-Son Dao, Guoxian Song, Tat-Jen Cham, and Jianfei Cai. Visiting the invisible: Layer-by-layer completed scene decomposition, 2021. 2

[66] Q. Zhou, S. Wang, Y. Wang, Z. Huang, and X. Wang. Human de-occlusion: Invisible perception and recovery for humans. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[67] Yan Zhu, Yuandong Tian, Dimitris Mexatas, and Piotr Dollár. Semantic amodal segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1