Amodal Completion via Progressive Mixed Context Diffusion
Supplementary Material

Katherine Xu[1] Lingzhi Zhang[2] Jianbo Shi[1]
[1]University of Pennsylvania, [2]Adobe Inc.
https://k8xu.github.io/amodal/

1. Implementation Details

1.1. Progressive Occlusion-aware Completion

To recover unoccluded appearances of objects, our method inpaints necessary regions by identifying occluders and iteratively performs this inpainting step to avoid incompletion.

Mask analysis. The first step of each iteration is performing instance segmentation [18, 29] and analyzing the object masks to determine occluders. Given the modal mask M_{modal} of a query object, we find its neighboring masks $\mathcal{M}_{neighbor}$. Then, we perform a depth ordering analysis and filter $\mathcal{M}_{neighbor}$ to contain masks closer to the camera than M_{modal}. This gives a refined set of masks, $\mathcal{M}_{occluder}$, which is aggregated into a single binary occlusion mask M_{occ}. Figure 1 shows additional examples of this mask analysis step.

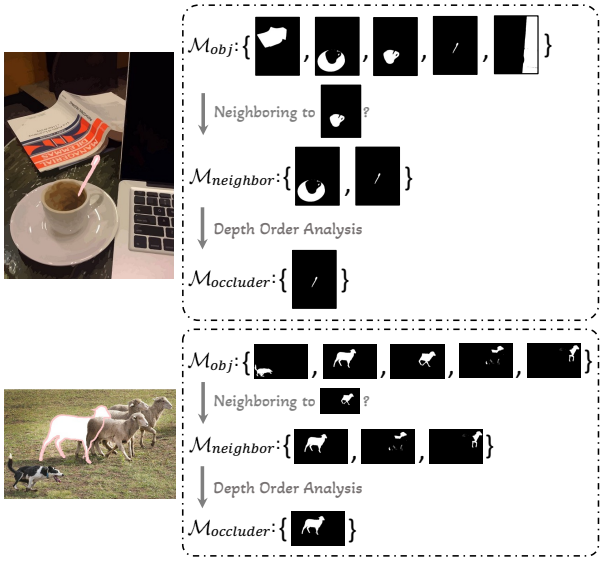

Figure 1. Mask analysis examples. **Top:** The cup is occluded by a spoon. **Bottom:** The sheep is occluded by another sheep.

Conditional padding. If the query object lies within 10 pixels of the image boundary, then we pad the image I_{in} and mask M_{occ} with white pixels in those directions.

Diffusion process and occlusion check. To create the input bundle, we crop the image I_{in} and mask M_{occ} into squares around the query object. First, we extend the object's tight bounding box by α pixels on each side. Second, if the object touches the image boundary, then we extend the bounding box by an additional β pixels on each side. Third, if the bounding box is not a square, then we extend the two shorter sides to have the same length as the longer sides. We use $\alpha = 60$, $\beta = 60$ in our experiments.

Moreover, we identify new occlusions by repeating the segment and mask analysis steps, as well as checking if the query object lies within λ pixels of the image boundary. We use $\lambda = 10$ to account for potentially inaccurate segmentations from the grounded segmentation models [18, 29] on the image boundary.

1.2. Mixed Context Diffusion Sampling

We describe the details of segmenting the query object in the noisy image $I^k_{syn\_amodal}$.

Extract features. At the k^{th} DDIM timestep, we extract text-conditioned features from the l^{th} UNet decoder layer. We experimentally determined that $l = 3$ is favorable for capturing an object's shape and $k = 20$ is generally better for achieving successful completions on difficult co-occurrence cases, but we acknowledge that the precise layer and timestep often depends on the occlusion.

Cluster features. The features from the third decoder layer are flattened and permuted from size $(640, 64, 64)$ to $(64 \times 64, 640)$. We perform unsupervised clustering [30] of the features and reshape the cluster assignments to $(64, 64)$.

Align clusters with query object. To segment the object in a noisy image, we upsample the cluster assignments to the size of the query object's modal mask, and then select the clusters with more than 20% overlap with the modal mask. Additionally, we include the modal mask in the segmentation, and we constrain the segmentation to only regions that are within the modal mask or occluder mask.

1.3. Counterfactual Completion Curation System

Our training-free rule outpaints generated objects and classifies them as complete or incomplete using two parameters: 1) the object's proximity to the image boundary in I'_{amodal}, and 2) the extension of the amodal mask area from M_{amodal} to M'_{amodal}. First, if any object pixel in M'_{amodal} is within γ pixels of the image boundary, then we judge the object as incomplete due to potential occlusion. Second, we dilate M_{amodal} using a 5×5 kernel and δ iterations. If M'_{amodal} is contained in the dilated M_{amodal}, then we judge the object as complete to allow minor extensions of complete objects. If $M'_{amodal} > \epsilon M_{amodal}$, then we judge the object as incomplete due to major extension. We experimentally determined $\gamma = 2, \delta = 4, \epsilon = 1.2$. Figure 2 shows additional examples of complete and incomplete objects determined by our rule.

2. Additional Qualitative Results

We use our method to complete objects in natural images. Figure 3 illustrates amodal completions within and beyond the image boundary. Figure 4 presents diverse versions of completed objects. Figure 5 visualizes non-cherry picked amodal completions of highly occluded objects in diverse images. Figure 6 compares our method with prior works. Figure 7 shows comparisons with Naive Outpainting and ours without Mixed Context Diffusion.

3. Experiment Details

3.1. User Studies

User preference study. For each study, we place the generated object from each method side by side, and then ask MTurk workers to select the object that looks most complete and realistic given the original image. Humans must correctly answer five out of six attention check images.

For comparisons with prior works, we asked at least three humans to label each batch of images, with a total of 110 images in four batches. For comparisons with Naive Outpainting and our method without Mixed Context Diffusion Sampling, we asked at least six humans to label each batch of images, with a total of 100 images in two batches.

User study for successful completions. We randomly select 100 occluded objects in natural images and generate their amodal completions using each method. We ask at least three MTurk workers to label each batch with around 50 images. Humans are given three choices: complete object, incomplete/overextended object, or discard if they are unsure. They must correctly answer four out of five attention check images. Figure 8 shows instructions displayed to users.

Human consensus for amodal completion. We asked three humans to judge the generated object in each amodal completion image as complete or incomplete. The human consensus is determined by a simple majority vote.

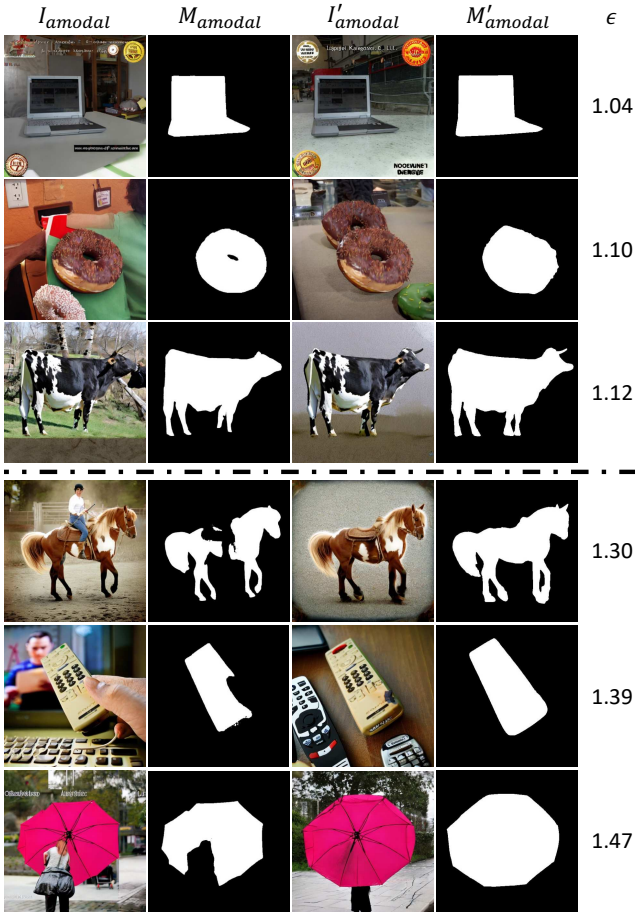

Figure 2. Complete and incomplete objects predicted by our counterfactual rule. If the new object mask M'_{amodal} is greater than $\epsilon = 1.2$ times the previous object mask M_{amodal}, then the object is predicted as incomplete. **Top:** Complete. **Bottom:** Incomplete.

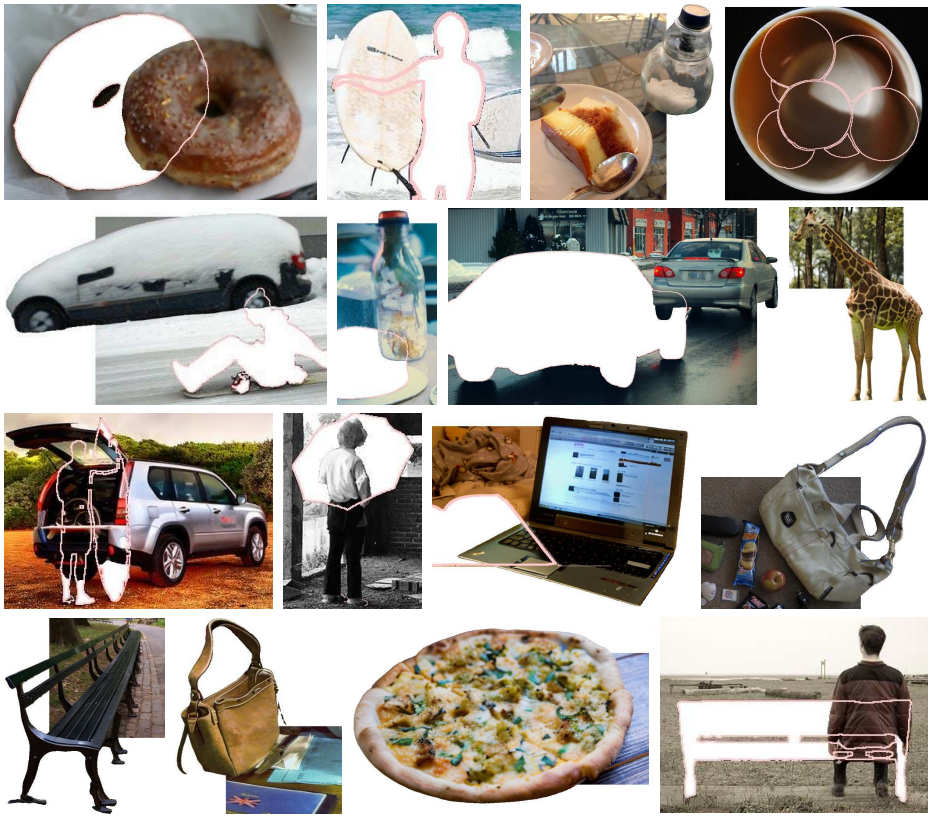

Figure 3. Our method completes objects within and beyond the image boundary.

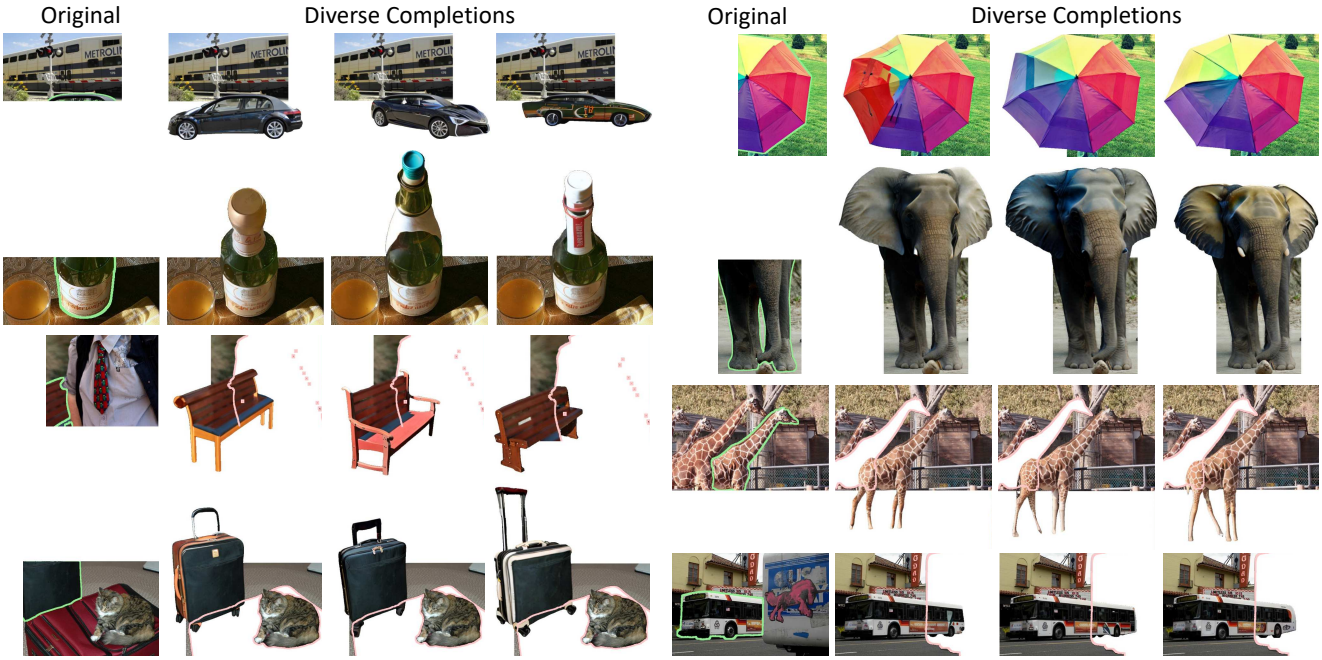

Figure 4. We can obtain diverse amodal completions for each occluded object.

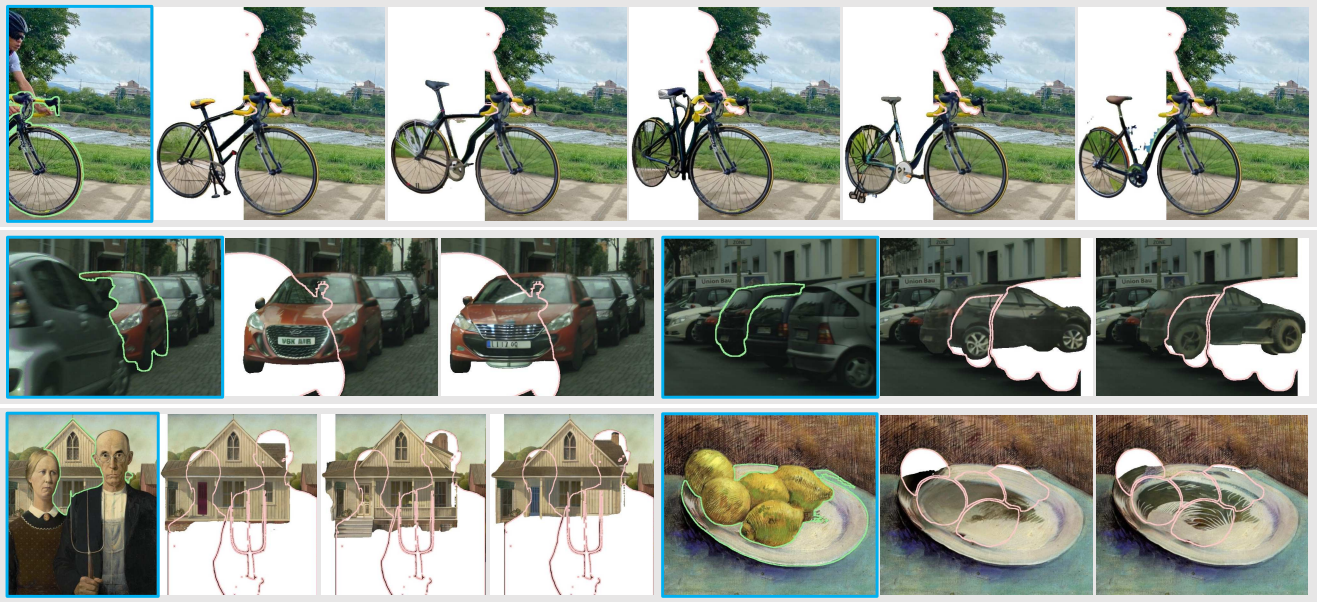

Figure 5. Non-cherry picked completions for highly occluded objects in diverse images. **Top:** Bicycle in an in-the-wild photo. **Middle:** Cars in the Cityscapes dataset. **Bottom:** House and plate in visual art.

Figure 6. Comparisons with prior works. **Top:** Easy cases. **Bottom:** Hard cases where the occluder is the top co-occurring object category for the query object. We find co-occurrence by randomly sampling 15,000 COCO images [26] and analyzing objects that appear together.

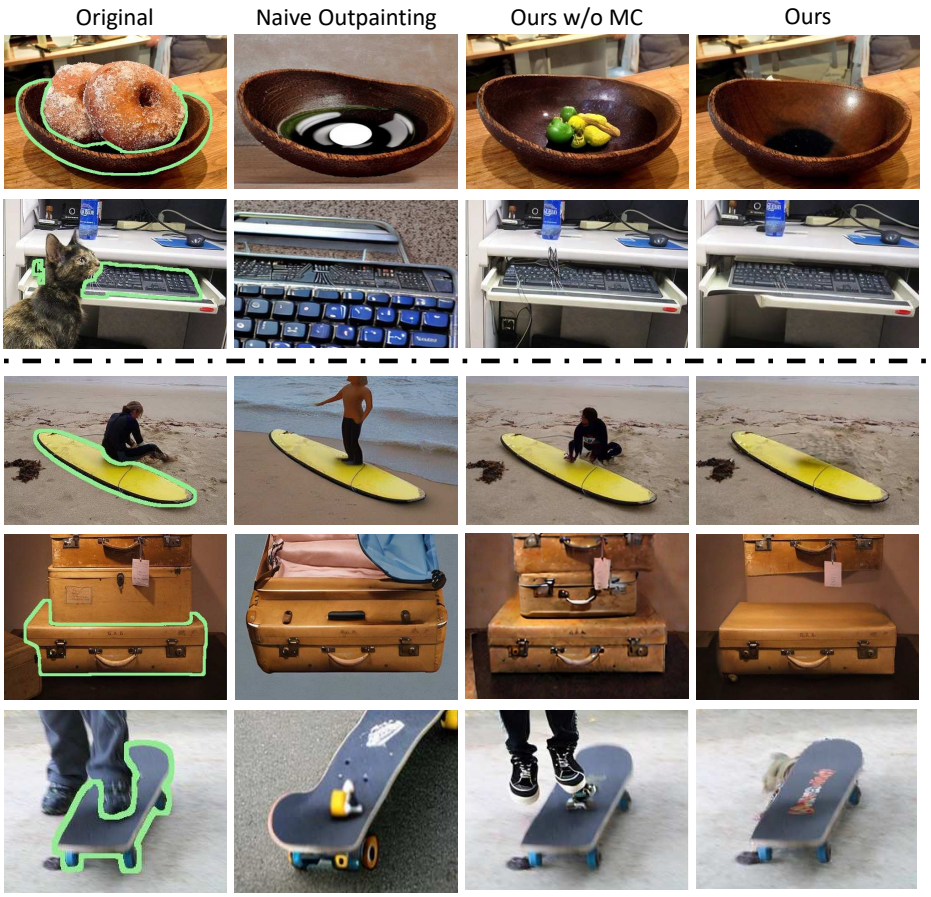

Figure 7. Comparisons with Naive Outpainting and our method without Mixed Context (MC) Diffusion Sampling. **Top:** Easy cases. **Bottom:** Hard cases where the occluder is the top co-occurring object category for the query object.

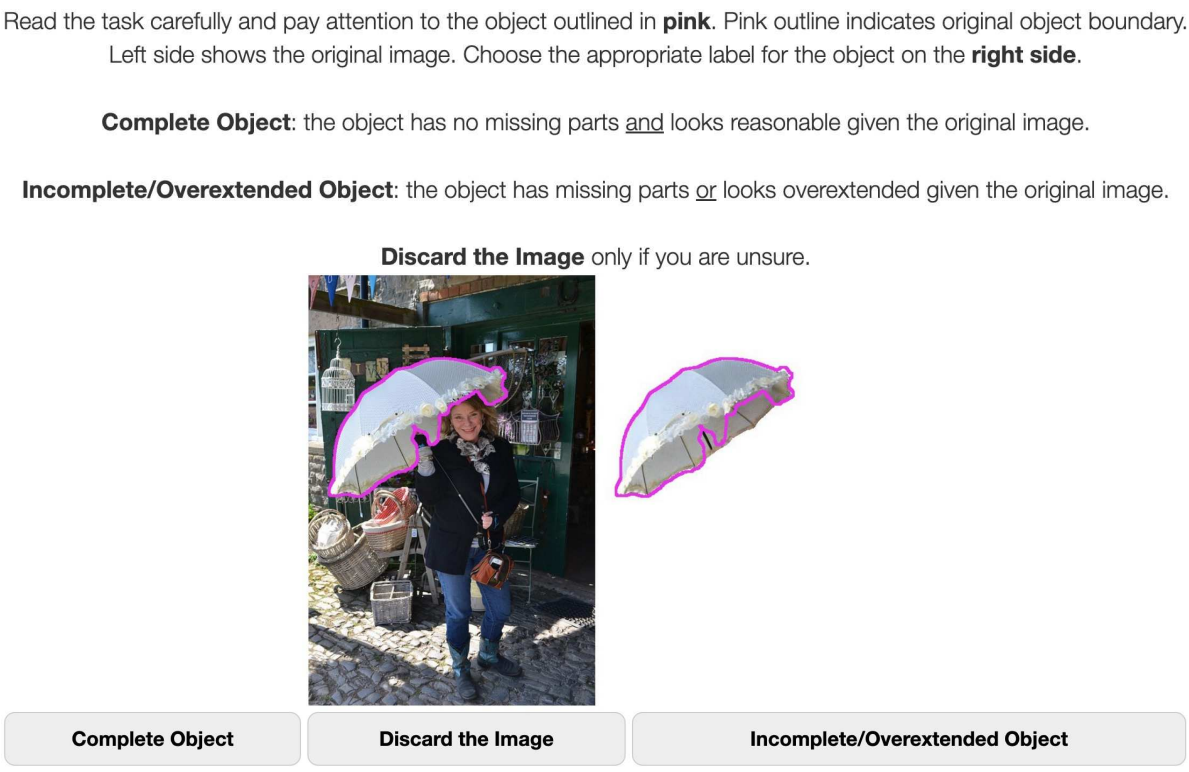

Figure 8. Example set of instructions displayed to MTurk workers for our user study to determine successful completions.