pix2gestalt: Amodal Segmentation by Synthesizing Wholes

Ege Ozguroglu[1] Ruoshi Liu[1] Dídac Surís[1] Dian Chen[2] Achal Dave[2] Pavel Tokmakov[2] Carl Vondrick[1]
[1]Columbia University [2]Toyota Research Institute
gestalt.cs.columbia.edu

Abstract

We introduce pix2gestalt, a framework for zero-shot amodal segmentation, which learns to estimate the shape and appearance of whole objects that are only partially visible behind occlusions. By capitalizing on large-scale diffusion models and transferring their representations to this task, we learn a conditional diffusion model for reconstructing whole objects in challenging zero-shot cases, including examples that break natural and physical priors, such as art. As training data, we use a synthetically curated dataset containing occluded objects paired with their whole counterparts. Experiments show that our approach outperforms supervised baselines on established benchmarks. Our model can furthermore be used to significantly improve the performance of existing object recognition and 3D reconstruction methods in the presence of occlusions.

1. Introduction

Although only parts of the objects in Figure 1 are visible, you are able to visualize the whole object, recognize the category, and imagine its 3D geometry. Amodal completion is the task of predicting the whole shape and appearance of objects that are not fully visible, and this ability is crucial for many downstream applications in vision, graphics, and robotics. Learned by children from an early age [30], the ability can be partly explained by experience, but we seem to be able to generalize to challenging situations that break natural priors and physical constraints with ease. In fact, we can imagine the appearance of objects during occlusions that cannot exist in the physical world, such as the horse in Magritte's *The Blank Signature*.

What makes amodal completion challenging compared to other synthesis tasks is that it requires grouping for both the visible and hidden parts of an object. To complete an object, we must be able to first recognize the object from partial observations, then synthesize only the missing regions for the object. Computer vision researchers and gestalt psychologists have extensively studied amodal completion in the past [10, 17, 18, 21, 33, 35, 49, 53], creating models that explicitly learn figure-ground separation. However, the prior work has been limited to representing objects in closed-world settings, restricted to only operating on the datasets on which they trained.

In this paper, we propose an approach for zero-shot amodal segmentation and reconstruction by learning to synthesize whole objects first. Our approach capitalizes on denoising diffusion models [14], which are excellent representations of the natural image manifold and capture all different types of whole objects and their occlusions. Due to their large-scale training data, we hypothesize such pretrained models have implicitly learned amodal representations (Figure 2), which we can reconfigure to encode object grouping and perform amodal completion. By learning from a synthetic dataset of occlusions and their whole counterparts, we create a conditional diffusion model that, given an RGB image and a point prompt, generates whole objects behind occlusions and other obstructions.

Our main result is showing that we are able to achieve state-of-the-art amodal segmentation results in a zero-shot setting, outperforming the methods that were specifically supervised on those benchmarks. We furthermore show that our method can be used as a drop-in module to significantly improve the performance of existing object recognition and 3D reconstruction methods in the presence of occlusions. An additional benefit of the diffusion framework is that it allows sampling several variations of the reconstruction, naturally handling the inherent ambiguity of the occlusions.

2. Related Work

We briefly review related work in amodal completion, analysis by synthesis, and denoising diffusion models for vision.

2.1. Amodal Completion and Segmentation

In this work, we define amodal completion as the task of generating the image of the whole object [10, 49], amodal segmentation as generating the segmentation mask of the whole object [18, 21, 33, 35, 53], and amodal detection as predicting the bounding box of the whole object [15, 17]. Most prior work focuses on the latter two tasks, due to the challenges in generating the (possibly ambiguous) pixels behind an occlusion. In addition, to our knowledge, all prior work on these tasks is limited to a small closed-world

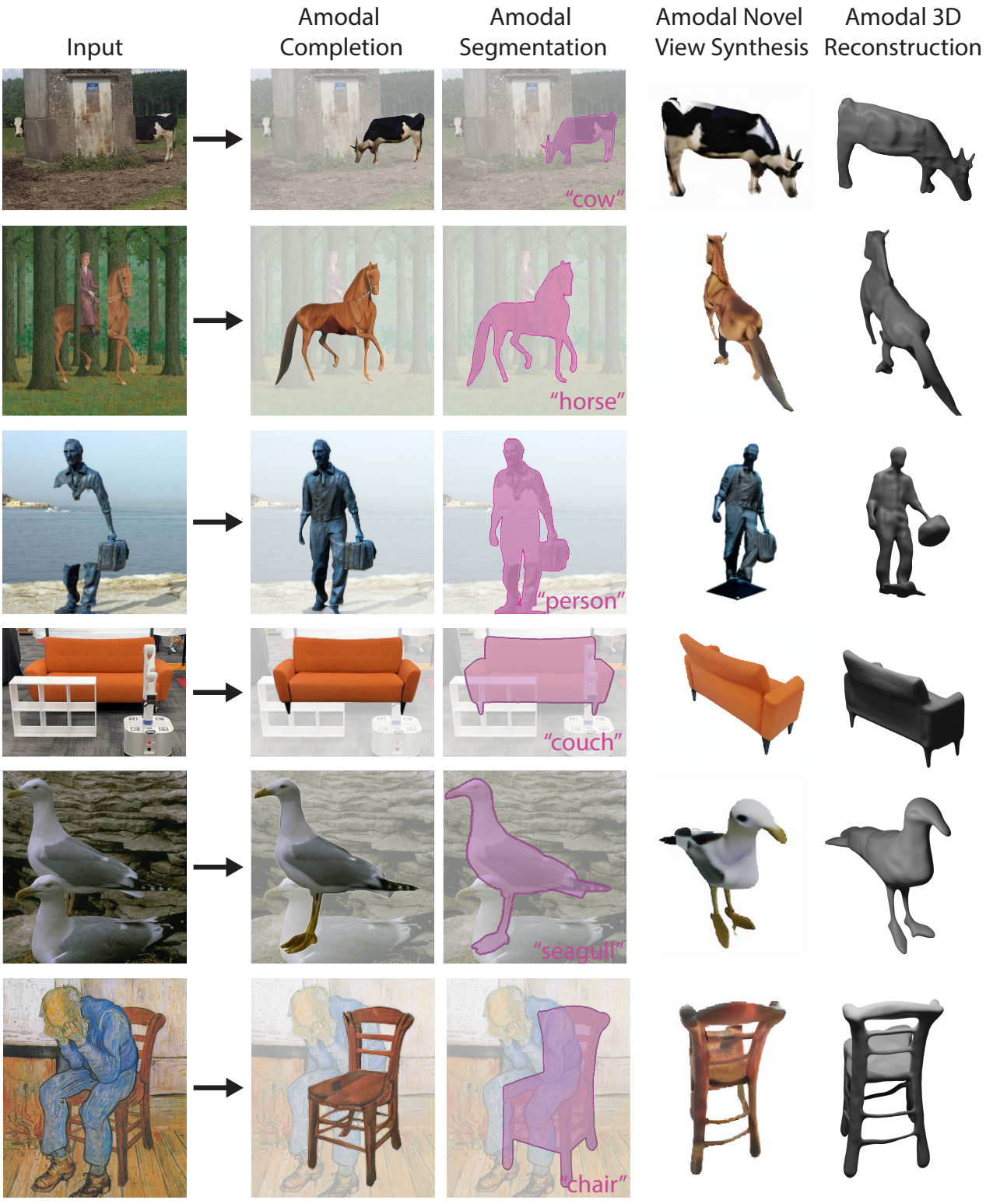

Figure 1. **Amodal Segmentation and Reconstruction via Synthesis.** We present ***pix2gestalt***, a method to synthesize whole objects from only partially visible ones, enabling amodal segmentation, recognition, novel-view synthesis, and 3D reconstruction of occluded objects.

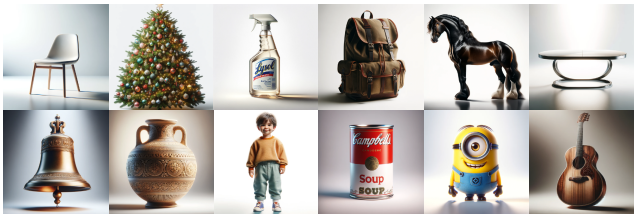

Figure 2. **Whole Objects**. Pre-trained diffusion models are able to generate all kinds of whole objects. We show samples conditioned on a category from Stable Diffusion. We leverage this synthesis ability for zero-shot amodal reconstruction and segmentation.

of objects [17, 18, 21, 33, 49] or to synthetic data [10]. For example, PCNet [49], the previous state-of-the-art method for amodal segmentation, operates only on a closed-world set of classes in Amodal COCO [53].

Our approach, by contrast, provides rich image completions with accurate masks, generalizing to diverse zero-shot settings, while still outperforming state-of-the-art methods in a closed-world. To achieve this degree of generalization, we capitalize on large-scale diffusion models, which implicitly learn internal representations of whole objects. We propose to unlock this capability by fine-tuning a diffusion model on a synthetically generated, realistic dataset of varied occlusions.

2.2. Analysis by Synthesis

Our approach is heavily inspired by analysis by synthesis [47] – a generative approach for visual reasoning. Image parsing [42] was a representative work that unifies segmentation, recognition, and detection by generation. Prior works have applied the analysis by synthesis approaches on various problems including face recognition [5, 42], pose estimation [27, 51], 3D reconstruction [22, 23], semantic image editing [1, 24, 52]. In this paper, we aim to harness the power of generative models trained with internet-scale data for the task of amodal completion, thereby aiding various tasks such as recognition, segmentation, and 3D reconstruction in the presence of occlusions.

2.3. Diffusion Models

Recently, Denoising Diffusion Probabilistic Model [14], or DDPM, has emerged as one of the most widely used generative architectures in computer vision due to its ability to model multi-modal distributions, training stability, and scalability. [8] first showed that diffusion models outperform GANs [12] in image synthesis. Stable Diffusion [36], trained on LAION-5B [39], applied diffusion model in the latent space of a variational autoencoder [19] to improve computational efficiency. Later, a series of major improvements were made to improve diffusion model performance [13, 41]. With the release of Stable Diffusion as a strong generative prior, many works have adapted it to solve tasks in different domain such as image editing [6, 11, 37], 3D [7, 25, 45], and modal segmentation [2, 3, 46]. In this work, we leverage the strong occlusion and complete object priors provided by internet-pretrained diffusion model to solve the zero-shot amodal completion task.

3. Amodal Completion via Generation

Given an RGB image x with an occluded object that is partially visible, our goal is to predict a new image with the shape and appearance of the whole object, and only the whole object. Our approach will accept any point or mask as a prompt p indicating the modal object:

$$\hat{x}_p = f_\theta(x, p)$$

where $\hat{x}_p$ is our estimate of the whole object indicated by p. Mapping from x to this unified whole form, *i.e. gestalt* of the occluded object, we name our method **pix2gestalt**. We want $\hat{x}$ to be perceptually similar to the true but unobserved whole of the object as if there was no occlusion. We will use a conditional diffusion model (see Figure 3) for f_θ.

An advantage of this approach is that, once we estimate an image of the whole object $\hat{x}$, we are able to perform any other computer vision task on it, providing a unified method to handle occlusions across different tasks. Since we will directly synthesize the pixels of the whole object, we can aid off-the-shelf approaches to perform segmentation, recognition, and 3D reconstruction of occluded objects.

To perform amodal completion, f needs to learn a representation of whole objects in the visual world. Due to their scale of training data, we will capitalize on large pretrained diffusion models, such as Stable Diffusion, which are excellent representations of the natural image manifold and have the support to generate unoccluded objects. However, although they generate high-quality images, their representations do not explicitly encode the grouping of objects and their boundaries to the background.

3.1. Whole-Part Pairs

To learn the conditional diffusion model f with the ability for grouping, we build a large-scale paired dataset of occluded objects and their whole counterparts. Unfortunately, collecting a natural image dataset of these pairs is challenging at scale. Prior datasets provide amodal segmentation annotations [33, 53], but they do not reveal the pixels behind an occlusion. Other datasets have relied on graphical simulation [16], which lack the realistic complexity and scale of everyday object categories.

We build paired data by automatically overlaying objects over natural images. The original images provide ground-truth for the content behind occlusions. However, we need to ensure that we only occlude whole objects in this construction, as otherwise our model could learn to generate

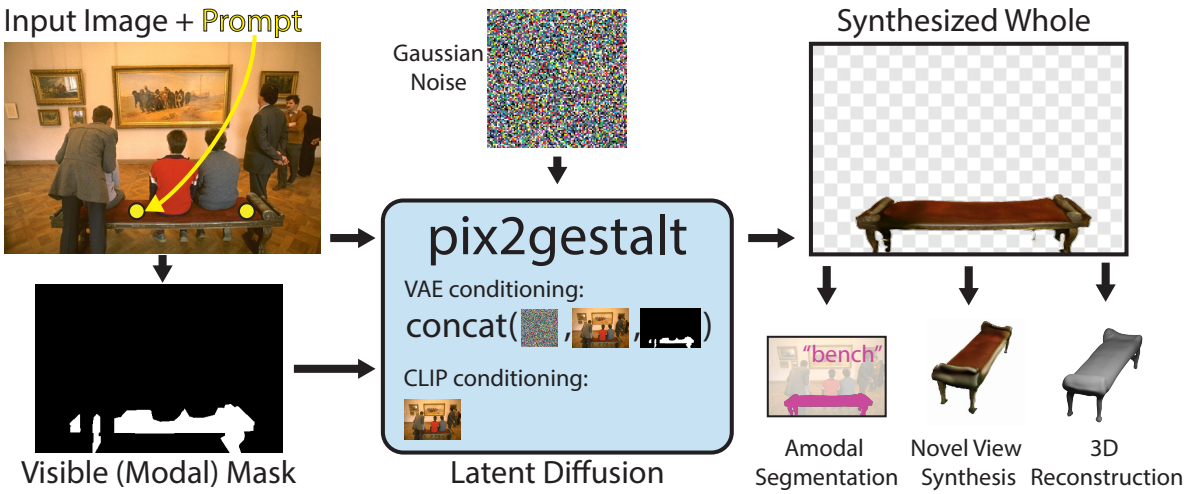

Figure 3. **pix2gestalt** is an amodal completion model using a latent diffusion architecture. Conditioned on an input occlusion image and a region of interest, the whole (amodal) form is synthesized, thereby allowing other visual tasks to be performed on it too. For conditioning details, see section 3.2.

incomplete objects. To this end, we use a heuristic that, if the object is closer to the camera than its neighboring objects, then it is likely a whole object. We use Segment Anything [20] to automatically find object candidates in the SA-1B dataset, and use the off-the-shelf monocular depth estimator MiDaS [4] to select which objects are whole. For each image with at least one whole object, we sample an occluder and superimpose it, resulting in a paired dataset of $837K$ images and their whole counterparts. Figure 4 illustrates this construction and shows examples of the heuristic.

3.2. Conditional Diffusion

Given pairs of an image x and its whole counterpart $\hat{x}_p$, we fine-tune a conditional diffusion model to perform amodal completion while maintaining the zero-shot capabilities of the pre-trained model. We solve for the following latent diffusion objective:

$$\min_{\theta} \mathbb{E}_{z\sim\mathcal{E}(x),t,\epsilon\sim\mathcal{N}(0,1)} \left[||\epsilon - \epsilon_\theta(z_t, \mathcal{E}(x), t, \mathcal{E}(p), \mathcal{C}(x))||_2^2\right]$$

where $0 \leq t < 1000$ is the diffusion time step, z_t is the embedding of the noised amodal target image $\hat{x}_p$. $\mathcal{C}(x)$ is the CLIP embedding of the input image, and $\mathcal{E}(\cdot)$ is a VAE embedding. Following [6, 25], we apply classifier-free guidance (CFG) [13] by setting the conditional information to a null vector randomly.

Amodal completion requires reasoning about the whole shape, its appearance, and contextual visual cues of the scene. We adapt the design in [6, 25] to condition the diffusion model ϵ_θ in two separate streams. $\mathcal{C}(x)$ conditions the diffusion model ϵ_θ via cross-attention on the semantic features of the partially visible object in x as specified by p, providing high-level perception. On the VAE stream, we channel concatenate $\mathcal{E}(x)$ and z_t, providing low-level visual details (shade, color, texture), as well as $\mathcal{E}(p)$ to indicate the visible region of the object.

After ϵ_θ is trained, f can generate $\hat{x}_p$ by performing iterative denoising [36]. The CFG can be scaled to control impact of the conditioning on the completion.

3.3. Amodal Base Representations

Since we synthesize RGB images of the whole object, our approach makes it straightforward to equip various computer vision methods with the ability to handle occlusions. We discuss a few common cases.

Image Segmentation aims to find the spatial boundaries of an object given an image x and an initial prompt p. We can perform amodal segmentation by completing an occluded object with f, then thresholding the result to obtain an amodal segmentation map. Note that this problem is under-constrained as there are multiple possible solutions. Given the uncertainty, we found that sampling multiple completions and performing a majority vote on the segmentation masks works best in practice.

Object Recognition is the task of classifying an object located in an bounding box or mask p. We can zero-shot recognize significantly occluded objects by first completing the whole object with f, then classifying the amodal completion with CLIP.

3D Reconstruction estimates the appearance and geometry of an object. We can zero-shot reconstruct objects with partial occlusions by first completing the whole object with f, then applying SyncDreamer and Score Distillation Sampling [32] to estimate a textured mesh.

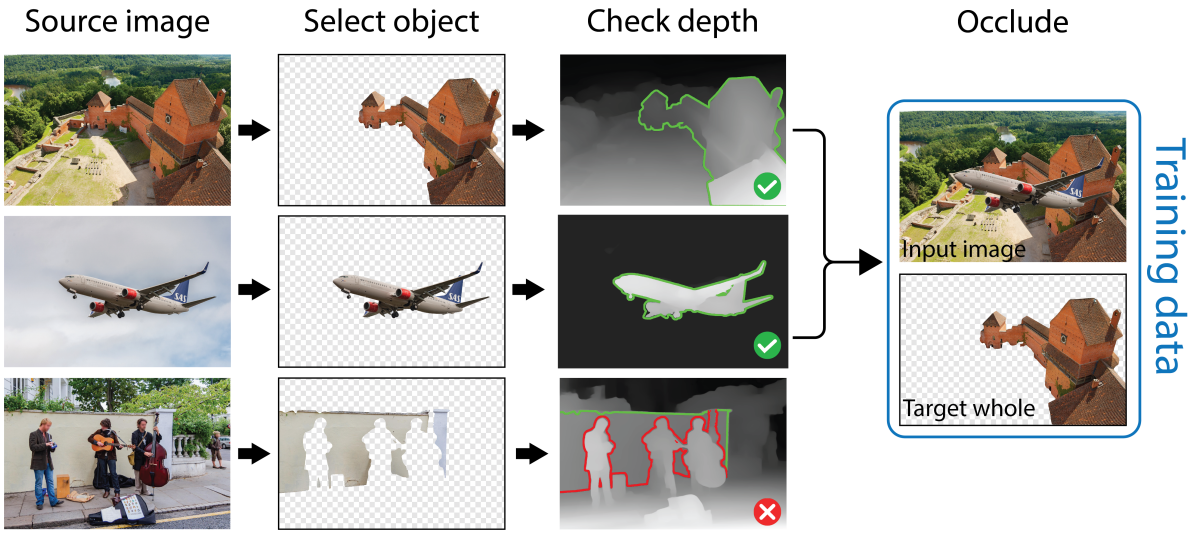

Figure 4. **Constructing Training Data.** To ensure we only occlude whole objects, we use a heuristic that objects closer to the camera than its neighbors are likely whole objects. The green outline around the object shows where the estimated depth is closer to the camera than the background (the red shows when it is not).

4. Experiments

We evaluate pix2gestalt's ability to perform zero-shot amodal completion for three tasks: amodal segmentation, occluded object recognition, and amodal 3D reconstruction. We show that our method provides amodal completions that directly lead to strong results in all tasks.

4.1. Amodal Segmentation

Setup. Amodal segmentation requires segmenting the full extent of a (possibly occluded) object. We evaluate this task on the Amodal COCO (COCO-A) [53] and Amodal Berkeley Segmentation (BSDS-A) datasets [28]. For evaluation, COCO-A provides 13,000 amodal annotations of objects in 2,500 images, while BSDS-A provides 650 objects from 200 images. For both datasets, we evaluate methods that take as input an image and a (modal) mask of the visible extent of an object, and output an amodal mask of the full-extent of the object. Following [49], we evaluate segmentations using mean intersection-over-union (mIoU). We follow the strategy in Section 3.3 to convert our amodal completions into segmentation masks.

We evaluate three baselines for amodal segmentation. The first method is PCNet [49], which is trained for amodal segmentation specifically for COCO-A. Next, we compare to two zero-shot methods, which do not train on COCO-A: Segment Anything (SAM) [20], a strong *modal* segmentation method, and Inpainting using Stable Diffusion-XL [31]. To evaluate inpainting methods, we provide as input an image with all but the visible object region erased, and convert the completed image output by the method into an amodal segmentation mask following the same strategy as for our method.

Results. Table 1 compares pix2gestalt with prior work. Despite never training on the COCO-A dataset, our method outperforms all baselines, including PCNet, which uses COCO-A images for training, and even PCNet-Sup, which is supervised using human-annotated amodal segmentations from COCO-A's training set. Compared to other zero-shot methods, our improvements are dramatic, validating the generalization abilities of our method. Notably, we also outperform the inpainting baseline which is based off a larger, more recent variant of Stable Diffusion [31]. This demonstrates that internet-scale training alone is not sufficient and our fine-tuning approach is key to reconfigure priors from pre-training for amodal completion.

We further analyze amodal completions qualitatively in Figure 6. While SD-XL often hallucinates extraneous, unrealistic details (e.g. person in front of the bus in the second row), PCNet tends to fail to recover the full extent of objects—often only generating the visible region, as in the Mario example in the third row. In contrast, pix2gestalt provides accurate, complete reconstructions of occluded objects on both COCO-A (Figure 6) and BSDS-A (Figure 7). Our method generalizes well beyond the typical occlusion scenarios found in those benchmarks. Figure 5 shows several examples of out-of-distribution images, including art pieces, illusions, and images taken by ourselves that are successfully handled by our method. Note that no prior work has shown open-world generalization (see 2.1).

Figure 8 illustrates the ability of the approach to generate diverse samples in shape and appearance when there is uncertainty in the final completion. For example, it is

Figure 5. **In-the-wild Amodal Completion and Segmentation.** We find that pix2gestalt is able to synthesize whole objects in novel situations, including artistic pieces, images taken by an iPhone, and illusions.

Table 1. **Amodal Segmentation Results**. We report **mIoU (%)** ↑ on Amodal COCO [53] and on Amodal Berkeley Segmentation Dataset [28, 53]. *PCNet-Sup trains using ground truth amodal masks from COCO-Amodal. See Section 4.1 for analysis.

Zero-shot	Method	COCO-A	BSDS-A
✗	PCNet [49]	81.35	-
✗	PCNet-Sup* [49]	82.53*	-
✓	SAM [20]	67.21	65.25
✓	SD-XL Inpainting [31]	76.52	74.19
✓	Ours	**82.87**	**80.76**
✓	Ours: Best of 3	**87.10**	**85.68**

able to synthesize several plausible completions of the occluded house in the painting. We quantitatively evaluate the diversity of our samples in the last row of Table 1 by sampling from our model three times and reporting the performance for the best sample ("Best of 3"). Finally, we found limitations of our approach in situations that require commonsense or physical reasoning. We show two examples in Figure 9.

4.2. Occluded Object Recognition

Next, we evaluate the utility of our method for recognizing occluded objects.

Setup. We use the Occluded and Separated COCO benchmarks [48] for evaluating classification accuracy under occlusions. The former consists of partially occluded objects, whereas Separated COCO contains objects whose modal region is separated into disjoint segments by the occluder(s), resulting in a more challenging problem setting. We evaluate on all 80 COCO semantic categories in the datasets using Top 1 and Top 3 accuracy.

We use CLIP [34] as the base open-vocabulary classifier. As baselines, we evaluate CLIP without any completion, reporting three variants: providing the entire image (CLIP), providing the entire image with a visual prompt (a red circle, as in Shtedritski *et al*. [40]) around the occluded object, or providing an image with all but the visible portion of the occluded object masked out. To evaluate our approach, we first utilize it to complete the occluded object, and then classify the output image using CLIP.

Results. Table 2 compares our method with the baselines. Visual prompting with a red circle (RC) and masking all but the visible object (Vis. Obj.) provide improvements over directly passing the image to CLIP on the simpler Occluded COCO benchmark, but fail to improve, and some times even decreases the performance of the baseline CLIP on the more challenging Separated COCO variant. Our method (Ours + CLIP), however, strongly outperforms all baselines for both the occluded and separated datasets, verifying the quality of our completions.

4.3. Amodal 3D Reconstruction

Finally, we evaluate our method for improving 3D reconstruction of occluded objects.

Setup. We focus on two tasks: Novel-view synthesis and

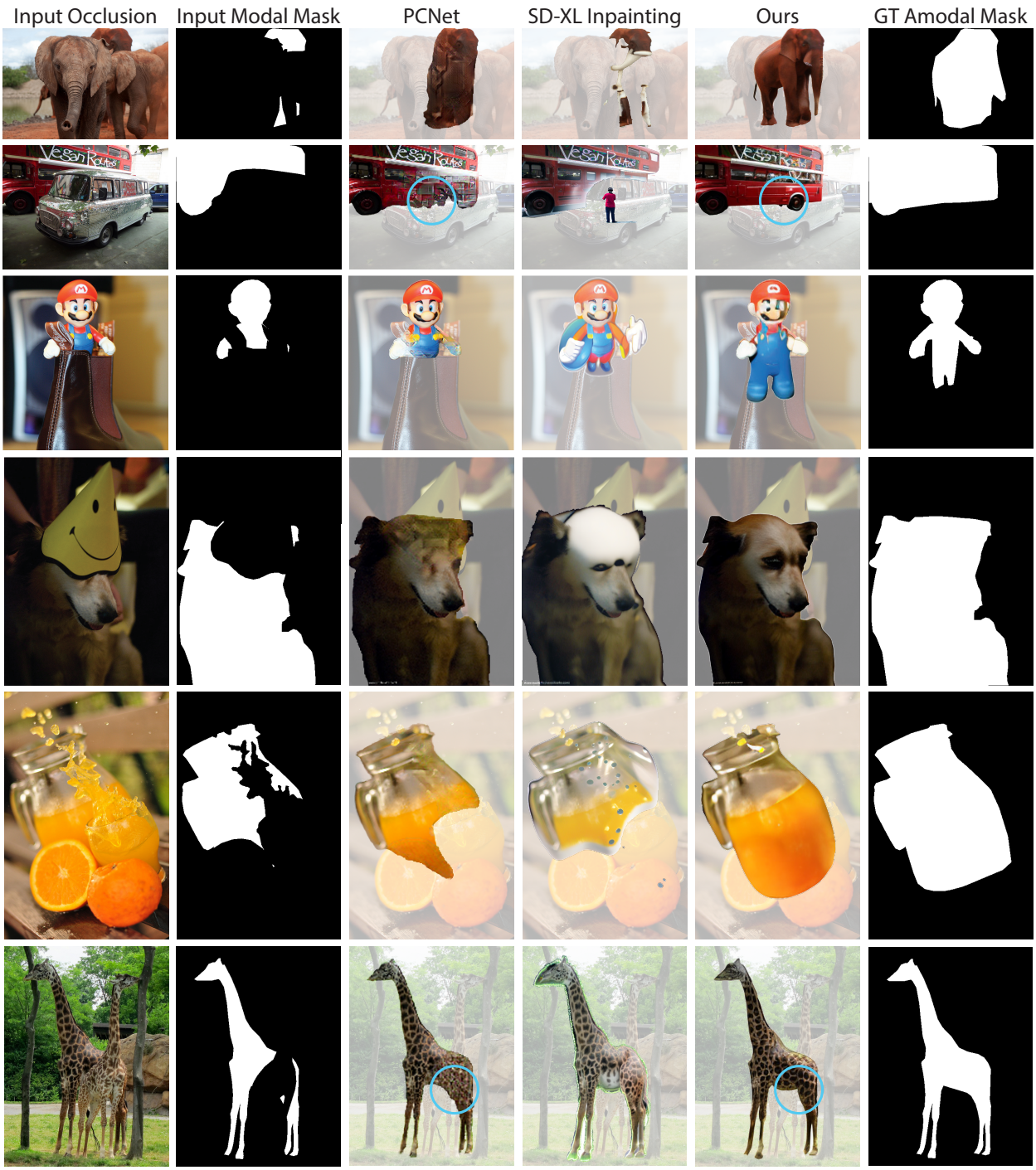

Figure 6. **Amodal Completion and Segmentation Qualitative Results on Amodal COCO**. In blue circles, we highlight completion regions that, upon a closer look, have a distorted texture in the PCNet baseline, and a correct one in our results.

single-view 3D reconstruction.

To demonstrate pix2gestalt's performance as a drop-in module to 3D foundation models [25, 26, 38], we replicate the evaluation procedure of Zero-1-to-3 [25, 26] on Google Scanned Objects (GSO) [9], a dataset of common household objects 3D scanned for use in embodied, synthetic, and 3D perception tasks. We use 30 randomly sampled objects from GSO ranging from daily objects to animals. For each object, we render a 256x256 image with synthetic occlusions sampled from the full dataset of 1,030 objects in GSO.

Figure 7. **Amodal Berkeley Segmentation Dataset Qualitative Results.** Our method provides accurate, complete reconstructions of occluded objects.

Figure 8. **Diversity in Samples**. Amodal completion has inherent uncertainties. By sampling from the diffusion process multiple times, the method synthesizes multiple plausible wholes that are consistent with the input observations.

Figure 9. **Common-sense and Physics Failures.** Left: reconstruction has the car going in the wrong direction. Right: reconstruction contradicts physics, failing to capture that a hand must be holding the donut box.

Table 2. **Occluded Object Recognition**. We report zero-shot classification accuracy on Occluded and Separated COCO [48]. While simple baselines fail to improve CLIP performance in the more challenging Separated COCO setting, our method consistently improves recognition accuracy by large margins. See Section 4.2 for analysis.

Method	Top 1 Acc. (%) ↑		Top 3 Acc. (%) ↑	
	Occluded	Sep.	Occluded	Sep.
CLIP [34]	23.33	26.04	43.84	43.19
CLIP + RC [40]	23.46	25.64	43.86	43.24
Vis. Obj. + CLIP	34.00	21.10	49.26	34.70
Ours + CLIP	**43.39**	**31.15**	**58.97**	**45.77**

We render from a randomly sampled view to avoid canonical poses, and generate two occluded images for each of the 30 objects, resulting in 60 samples.

For amodal novel-view synthesis, we quantitatively evaluate our method using 3 metrics: PSNR, SSIM [44], and LPIPS [50], measuring the image-similarity of the input and ground truth views. For 3D reconstruction, we use the Volumetric IoU and Chamfer Distance metrics. We compare our approach with SyncDreamer [26], a 3D generative model that fine-tunes Zero123-XL [7, 25] for multi-view consistent novel view synthesis and consequent 3D reconstruction with NeuS [43] and NeRF [29]. Our first baseline provides as input to SyncDreamer the segmentation mask of all foreground objects, following the standard protocol. To avoid reconstructing occluded objects, we additionally evaluate two variants that use SAM [20] to estimate the mask of only the object of interest, or the ground truth mask for the object of interest (GT Mask). Finally, to evaluate our method, we provide as input the full object completed by our method, along with the corresponding amodal mask. We evaluate two variants of our method: One where we provide a modal mask for the object of interested as estimated by SAM (Ours (SAM Mask)) and one where we use the ground truth modal mask (Ours (GT Mask)).

Results. We compare our approach with the two baselines in Table 4 for novel view synthesis and Table 3 for 3D reconstruction. Quantitative results demonstrate that we strongly outperform the baselines for both tasks. In novel-

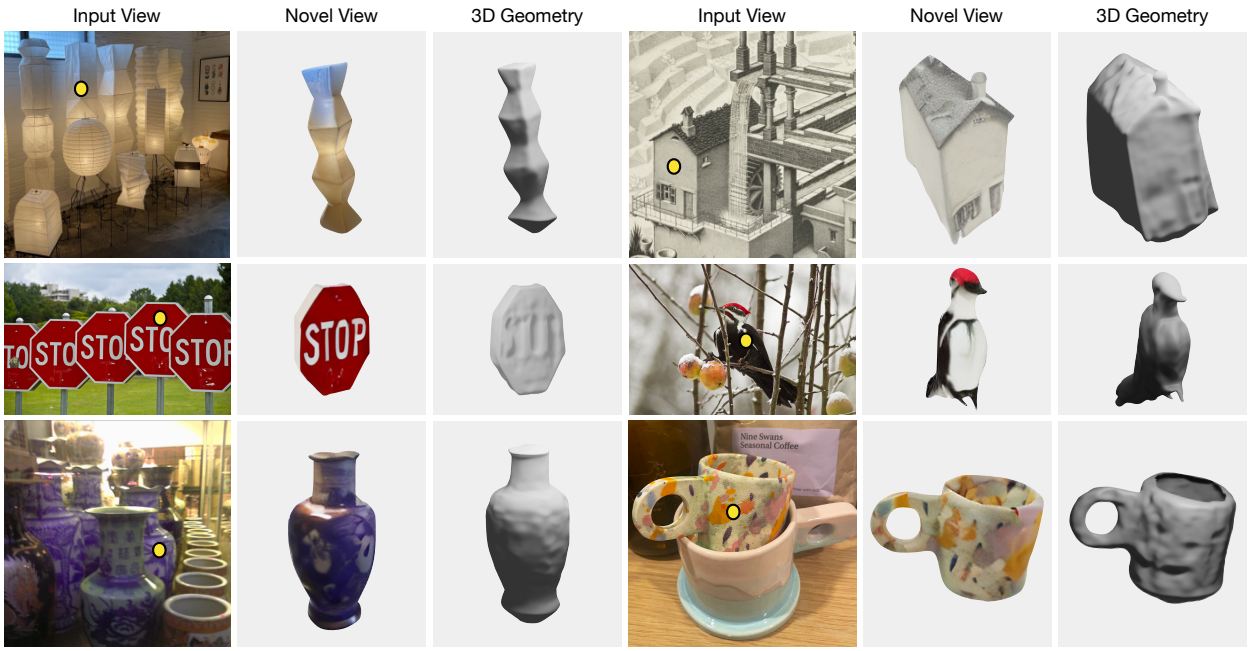

Figure 10. **Amodal 3D Reconstruction qualitative results**. The object of interest is specified by a point prompt, shown in yellow. Incorporating pix2gestalt as a drop-in module to state-of-the-art 3D reconstruction models allows us to address challenging and diverse occlusion scenarios with ease.

Table 3. **Single-view 3D Reconstruction.** We report Chamfer Distance and Volumetric IoU for Google Scanned Objects. See Section 4.3 for analysis.

	CD ↓	IoU ↑
SyncDreamer [26]	0.0884	0.2741
SAM Mask + SyncDr.	0.1182	0.0926
Ours (SAM Mask) + SyncDr.	**0.0784**	**0.3312**
GT Mask + SyncDr.	0.1084	0.1027
Ours (GT Mask) + SyncDr.	**0.0681**	**0.3639**

Table 4. **Novel-view synthesis from one image.** We report results on Google Scanned Objects [9]. Note SSIM measures image quality, not novel-view accuracy. See Section 4.3 for analysis.

	LPIPS ↓	PSNR ↑	SSIM ↑
SyncDreamer [26]	0.3221	11.914	0.6808
SAM + SyncDr.	0.3060	12.432	0.7248
Ours (SAM Mask) + SyncDr.	**0.2848**	**13.868**	0.7211
GT Mask + SyncDr.	0.2905	12.561	0.7322
Ours (GT Mask) + SyncDr.	**0.2631**	**14.657**	0.7328

view synthesis, we outperform SAM + SyncDreamer on the image reconstruction metrics, LPIPS [50] and PSNR [44]. Compared to SAM as a modal pre-processor, we obtain these improvements as a drop-in module to SyncDreamer while still retaining equivalent image quality (Table 4, SSIM [44]). With ground truth mask inputs, we obtain further image reconstruction gains. Moreover, even though our approach utilizes an additional diffusion step compared to SyncDreamer only, we demonstrate less image quality degradation.

For reconstruction of the 3D geometry, our fully automatic method outperforms all of the baselines for both volumetric IoU and Chamfer distance metrics, even the baselines that use ground masks. Providing the ground truth to our approach further improves the results. Figure 10 shows qualitative evaluation for 3D reconstruction of occluded objects, ranging from an Escher lithograph to in-the-wild images.

5. Conclusion

In this work, we proposed a novel approach for zero-shot amodal segmentation via synthesis. Our model capitalizes on whole object priors learned by internet-scale diffusion models and unlocks them via fine-tuning on a synthetically generated dataset of realistic occlusions. We then demonstrated that synthesizing the whole object makes it straightforward to equip various computer vision methods with the ability to handle occlusions. In particular, we reported state-of-the art results on several benchmarks for amodal segmen-

tation, occluded object recognition and 3D reconstruction.

Acknowledgements: This research is based on work partially supported by the Toyota Research Institute, the DARPA MCS program under Federal Agreement No. N660011924032, the NSF NRI Award #1925157, and the NSF AI Institute for Artificial and Natural Intelligence Award #2229929. DS is supported by the Microsoft PhD Fellowship.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN: How to embed images into the stylegan latent space? In *ICCV*, 2019. 3

[2] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 3

[3] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 3

[4] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 4

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. 2023. 3

[6] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 3, 4

[7] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 3, 8

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021. 3

[9] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3D scanned household items. In *ICRA*, 2022. 7, 9

[10] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *CVPR*, 2018. 1, 3

[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014. 3

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3, 4

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33, 2020. 1, 3

[15] Cheng-Yen Hsieh, Tarasha Khurana, Achal Dave, and Deva Ramanan. Tracking any object amodally, 2023. 1

[16] Y.-T. Hu, H.-S. Chen, K. Hui, J.-B. Huang, and A. G. Schwing. SAIL-VOS: Semantic Amodal Instance Level Video Object Segmentation – A Synthetic Dataset and Baselines. In *Proc. CVPR*, 2019. 3

[17] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Amodal completion and size constancy in natural scenes. In *ICCV*, 2015. 1, 3

[18] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *CVPR*, 2021. 1, 3

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[20] 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. In *ICCV*, 2023. 4, 5, 6, 8

[21] Huan Ling, David Acuna, Karsten Kreis, Seung Wook Kim, and Sanja Fidler. Variational amodal object completion. *NeurIPS*, 2020. 1, 3

[22] Ruoshi Liu and Carl Vondrick. Humans as light bulbs: 3d human reconstruction from thermal reflection. In *CVPR*, 2023. 3

[23] Ruoshi Liu, Sachit Menon, Chengzhi Mao, Dennis Park, Simon Stent, and Carl Vondrick. Shadows shed light on 3d objects. *arXiv preprint arXiv:2206.08990*, 2022. 3

[24] Ruoshi Liu, Chengzhi Mao, Purva Tendulkar, Hao Wang, and Carl Vondrick. Landscape learning for neural network inversion. In *ICCV*, 2023. 3

[25] 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. In *ICCV*, 2023. 3, 4, 7, 8

[26] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Learning to generate multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 7, 8, 9

[27] Wufei Ma, Angtian Wang, Alan Yuille, and Adam Kortylewski. Robust category-level 6D pose estimation with coarse-to-fine rendering of neural features. In *ECCV*, 2022. 3

[28] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001. 5, 6

[29] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 8

[30] Jean Piaget. *The construction of reality in the child*. Routledge, 2013. 1

[31] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and

Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 5, 6

[32] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 4

[33] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with KINS dataset. In *CVPR*, 2019. 1, 3

[34] 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, 8

[35] N Dinesh Reddy, Robert Tamburo, and Srinivasa G Narasimhan. Walt: Watch and learn 2d amodal representation from time-lapse imagery. In *CVPR*, 2022. 1

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3, 4

[37] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 3

[38] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*, 2023. 7

[39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5B: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022. 3

[40] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *ICCV*, 2023. 6, 8

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

[42] Zhuowen Tu, Xiangrong Chen, Alan L Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of computer vision*, 63:113–140, 2005. 3

[43] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 8

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 8, 9

[45] Rundi Wu, Ruoshi Liu, Carl Vondrick, and Changxi Zheng. Sin3dm: Learning a diffusion model from a single 3d textured shape. *arXiv preprint arXiv:2305.15399*, 2023. 3

[46] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. *arXiv preprint arXiv:2303.04803*, 2023. 3

[47] Alan Yuille and Daniel Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308, 2006. 3

[48] Guanqi Zhan, Weidi Xie, and Andrew Zisserman. A tri-layer plugin to improve occluded detection. *BMVC*, 2022. 6, 8

[49] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene deocclusion. In *CVPR*, 2020. 1, 3, 5, 6

[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 8, 9

[51] Yi Zhang, Pengliang Ji, Angtian Wang, Jieru Mei, Adam Kortylewski, and Alan Yuille. 3D-Aware neural body fitting for occlusion robust 3d human pose estimation. In *ICCV*, 2023. 3

[52] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain GAN inversion for real image editing. In *ECCV*, 2020. 3

[53] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *CVPR*, 2017. 1, 3, 5, 6