Open-vocabulary Self-Supervised Scene De-Occlusion

Zhangyu Lai[1], Xin Chen[1*], Yilin Lu[1*], Shiyuan Ma[1*], Wenbin Lai[1*]

[1]School of Informatics, Xiamen University
[2]Key Laboratory of Multimedia Trusted Perception and Efficient Computing
{31520231154295, 23020231154134, 23020231154212, 31520231154269, 23020231154143}@stu.xmu.edu.cn

Abstract

Natural scene understanding is a challenging task, particularly when confronted with images that involve occlusions of multiple object parts. While existing de-occlusion methods have achieved some success in closed-class domains, they still encounter significant challenges when applied to more intricate open-world scenarios, where there are unrestricted image domains and classes. In this paper, we aim to address the challenges associated with constructing training instances for existing occlusion handling methods. Furthermore, we propose the novel utilization of a diffusion model to tackle the de-occlusion problem, specifically focusing on open-world scenarios. Extensive experiments on real-world scenes demonstrate the superior performance of our approach to other alternatives. Our envisioned framework holds the potential to effectively handle de-occlusion in complex open-world scenes, consequently enhancing scene understanding capabilities for downstream tasks. By addressing these issues, we can provide more accurate input data for the currently popular 2D and 3D reconstruction tasks, thereby enhancing the quality and accuracy of the generated results.

(a) Image after random masking processing

(b) Generate images after model completion

Figure 1: Complete image effect display

Introduction

In our everyday life, we often observe partially occluded objects. Humans can reliably recognize the visible parts of an object and use them as cues to estimate the occluded parts. This perception of the object's complete structure under occlusion is referred to as amodal perception(Nanay 2018).

A key problem in amodal perception is scene deocclusion, which originates from the image Amodal completion and involves the subtasks of recovering the underlying occlusion ordering and completing the invisible parts of occluded objects. Existing computer vision systems can compete with humans in understanding the visible parts of objects, but still fall far short of humans when it comes to depicting the invisible parts of partially occluded objects(Zhan et al. 2020).

In many computer vision tasks, scene deocclusion is important to study, which able to acquire a full decomposition of a scene, with only an image as input, which conduces to a lot of applications, e.g. object-level image editing. In particular, autonomous vision systems applied in reality must perform similar reasoning for occluded objects to guarantee operational safety and reliability. For robotic grasping systems, the ability to infer the entire structure of an occluded object from its visible parts allows the robot to directly grasp and manipulate unseen objects in cluttered scenes. For self-driving, the rapid identification of an object and its complete spatial extent from local areas of the object in a complex scene helps predict more accurately what is likely to happen in the short future and thus plan accordingly. In addition, the scene deocclusion task is a very important upstream work for both 2D and 3D reconstruction. Through deocclusion, more abundant entity semantic information can be obtained, which can smooth obstacles to reconstruction.

Existing scene understanding systems mainly focus on recognizing the visible parts of a scene, ignoring the intact appearance of physical objects in the real world. To decompose a scene into instances with completed appearances is extremely challenging. This is be cause realistic natural scenes often consist of a vast collection of physical objects, with complex scene structure and occlusion relationships, especially when one object is occluded by multiple objects, or when instances have deep hierarchical occlusion relationships. Another challenge in this novel task is the lack of data: there is no complex, realistic dataset that provides

*These authors contributed equally.

intact ground-truth appearance for originally occluded objects and backgrounds in a scene. The current approaches is the requirement of detailed supervision of amodal object masks either through human annotation(Patrick Follmann and Ttger IEEE, 2019; Li and Malik Springer,2016; Lu Qi and Jia June 2019) or by generating artificially occluded images(Xiaohang Zhan and Loy June 2020).

In this work, we introduce a self-supervised scene deocclusion approach based on the diffusion model. Our proposed method addresses the significant disparity between training and testing data distributions encountered by Deocclusion(Xiaohang Zhan and Loy June 2020). We present a novel paradigm for constructing self-supervised occlusion removal data, training our model on the latest high-quality open-world scene dataset, Entityseg. Our approach departs from the conventional multi-stage occlusion recovery by aligning more closely with human visual perception.

Our experiments used the most commonly used amodal segmentation dataset: COCO Amodal(Lu Qi and Jia June 2019), showing comparable results to former approaches on datasets of real scenes. In summary, we make several contributions:

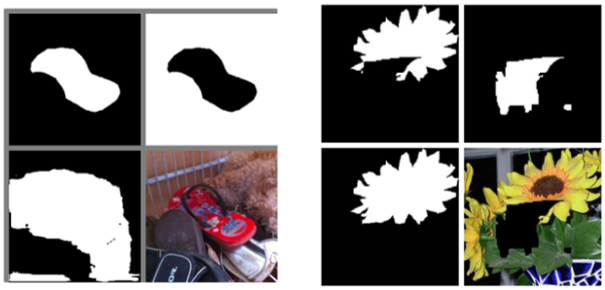
- We present a novel paradigm for constructing self-supervised occlusion removal data;
- We propose the innovative application of a diffusion model framework to tackle de-occlusion tasks;
- We validate the feasibility of this method on the open-world dataset Entityseg;
- Our scene deocclusion framework has the ability as an upstream task, to bring better results for 2D and 3D reconstruction tasks.

Related work

EntitySeg Dataset Numerous datasets have been proposed for semantic, instance, and panoptic segmentation, e.g., Microsoft COCO(Lin et al. 2014), ADE20K(Zhou et al. 2017), KITTI(Geiger, Lenz, and Urtasun 2012), PASCAL-VOC (Everingham et al. 2010), OpenImages(Kuznetsova et al. 2020), and so on. In addition, there are some datasets specially designed for specific scenarios, such as amodal segmentation (COCO-Amodal(Zhu et al. 2017), KINS(Qi et al. 2019)), human segmentation (CelebAMask-HQ(Lee et al. 2020), LIP(Gong et al. 2017), MHP(Zhao et al. 2018)) and domain adaptation (Synscapes(Wrenninge and Unger 2018), GTA5(Richter et al. 2016)). Despite the significant contributions from these datasets, there is still a need to fulfill the requirements of real-world applications with high-quality images of large diversity.The EntitySeg dataset used in this study consists of 33,227 images with high-quality mask annotations. Compared with existing datasets, it has three distinctive features. Firstly, 71.25% and 86.23% of the images are of high quality, with at least 2000px×2000px and 1000px×1000px, respectively, staying in line with the current digital imaging trends. Secondly, our dataset is open-world and not restricted to predefined classes. We consider each semantically-coherent region in the images as an entity, even if it is blurred or hard to recognize semantically. Thirdly, the mask annotation along the boundaries is more precise than those in other datasets.

Amodel completion Amodal completion is slightly different from amodal instance segmentation. In amodal completion, modal masks are given at test time and the task is to complete the modal masks into amodal masks. Previous works on amodal completion typically rely on heuristic assumptions on the invisible boundaries to perform amodal completion with given ordering relationships. Kimia et al. (Kimia, Frankel, and Popescu 2003) propose to adopt Euler Spiral in amodal completion. Lin et al.(Lin et al. 2016) use cubic B ezier curves. Silberman et al.(Silberman et al. 2014)apply curve primitives including straight lines and parabolas. Since these studies still require ordering as the input, they cannot be adopted directly to solve de-occlusion problem. Besides, these unsupervised approaches mainly focus on toy examples with simple shapes. Kar et al.(Kar et al. 2015) use keypoint annotations to align 3D object templates to 2D image objects, so as to generate the ground truth of amodal bounding boxes. Ehsani et al. (Ehsani, Mottaghi, and Farhadi 2018)leverage 3D synthetic data to train an end-to-end amodal completion network. Similar to unsupervised methods, our framework does not need annotations of amodal masks or any kind of 3D/synthetic data. In contrast, our approach, Self-Supervised Scene De-occlusion, can address the challenge of complete scene amodal completion in highly cluttered natural scenes, a task that other unsupervised methods are unable to handle.

Controlling Image Diffusion Models Controlling Image Diffusion Models facilitate personalization, customization, or task-specific image generation. The image diffusion process directly provides some control over color variation (Meng et al. 2021) and inpainting . Text-guided control methods focus on adjusting prompts, manipulating CLIP features, and modifying cross-attention (Avrahami, Lischinski, and Fried 2022; Brooks, Holynski, and Efros 2022; Gafni et al.). MakeAScene (Gafni et al.) encodes segmentation masks into tokens to control image generation. SpaText (Avrahami et al. 2022) maps segmentation masks into localized token embeddings. GLIGEN (Li et al. 2023) learns new parameters in attention layers of diffusion models for grounded generating. Textual Inversion and DreamBooth (Ruiz et al. 2022) can personalize content in the generated image by finetuning the image diffusion model using a small set of user-provided example images. Prompt-based image editing (Brooks, Holynski, and Efros 2022) provides practical tools to manipulate images with prompts. Voynov et al. (Voynov, Aberman, and Cohen-Or 2022) propose an optimization method that fits the diffusion process with sketches. Concurrent works (Omer et al. 2023) examine a wide variety of ways to control diffusion models.The ControlNet used in this study is a neural network architecture that enhances large pretrained text-to-image diffusion models through spatial localization and task-specific image conditions. It facilitates a broader range of applications for controlling image diffusion models.



(a) training stage (b) inference stage

Figure 2: The problem of distribution differences in the original method

Method

Train

Overview. As described in algorithm 1, our approach begins with the EntitySeg dataset, tailored for high-quality image segmentation. To effectively address occlusion removal in images, we propose a network architecture that combines EntitySeg and the Diffusion Model, which we refer to as MaskRemoveNet. This network leverages the CropFormer model for image segmentation, merging global and local details to achieve precise image segmentation.

MaskRemoveNet uses CropFormer to create image masks, providing high-quality training data. We then preprocess the input, extract conditions, and use them to enhance the performance of the Conditional Latent Diffusion Module.

This module, comprising a Stable Diffusion module for inpainting and a Control Net branch for specifying regions, achieves superior results in revealing obscured content. Our method combines these components to address the challenge of occlusion removal in high-resolution images.

Given an input image z_o, image diffusion algorithms progressively add noise to the image and produce a noisy image z_t, where t represents the number of times noise is added. Given a set of conditions including time step t, global condition c_g, as well as a local condition c_l, image diffusion algorithms learn a network ϵ_θ to predict the noise added to the noisy image z_t with:

$$L = E_{z_0,t,c_g,c_f,\epsilon\ N(0,1)[||\epsilon-\epsilon_\theta(z_t,t,c_g,c_l)||^2]} \quad (1)$$

where is the overall learning objective of the entire diffusion model. This learning objective is directly used in fine-tuning diffusion models with ControlNet.

EntitySeg (Qi et al. 2023) introduces a dataset called EntitySeg and employs a model named CropFormer to achieve high-quality image segmentation tasks. The EntitySeg dataset comprises high-resolution and high-quality images along with pixel-level mask annotations, characterized by open-world and non-predefined entity annotations as well as precise boundary annotations, which are particularly well-suited for our task of removing occlusions. CropFormer is a multi-view fusion method based on Transformer, which can effectively utilize global context information and local detail information to achieve high-quality image segmentation. Utilizing CropFormer to segment images and generate masks provides us with high-quality training data for our task.

CropFormer. Given the high-quality and high-resolution characteristics of the EntitySeg dataset, CropFormer was introduced by EntitySeg to address the challenging problem of instance-level segmentation on high-resolution images. It enhances mask prediction by fusing high-resolution image crops that provide finer-grained image details with complete images. MaskRemoveNet employs CropFormer to segment images and obtain masks, which are used to construct the dataset for subsequent training of the Conditional Latent Diffusion Module.

Pre-Process. The MaskRemoveNet receives an image with a mask as input. To generate the image with a mask, we add a random mask that has been aligned to the image to a random image to attain the input. The input is later resized and removed occlusion in the pre-processing block to extract global and local conditions which both feed to the Conditional Latent Diffusion Module for better performance. Meanwhile, the pre-processing module outputs the Ground Truth as the input of the Latent Encoder.

Condition Latent Diffusion Model. Disregarding the entirety of information may lead to the omission of crucial details that could otherwise contribute to a better understanding of the problem or task at hand. This can result in decision-making and solutions that are less comprehensive or accurate. To address this issue, we simultaneously employ both global and local information as conditioning factors in the Condition Latent Diffusion Model to govern its inpainting process. This approach yields superior results in uncovering obscured content.

The training process of the Stable Diffusion model involves using the frozen latent representation output from the Latent Encoder as input, denoted as x_0. Subsequently, noise z is incrementally added to the sequence from x0 to x_t. These steps are designed to optimize the performance of the Stable Diffusion model. During the training process, we utilize the noise $\hat{z}$ generated by the Stable Diffusion model as the predicted noise and calculate the loss between $\hat{z}$ and the actual noise z. Ultimately, by subtracting $\hat{z}$ from x_t, we obtain the image restored by the Stable Diffusion model.

To better understand this process, let's briefly summarize: starting from x_0, we progressively generate an image sequence by introducing noise z. Then, by comparing the generated noise $\hat{z}$ with the actual noise z, we guide the model to learn more accurate representations. Finally, by removing the noise generated by the model from the final image, we obtain the image reconstructed by the Stable Diffusion model.

The key objective of this training process is to enhance the model's capabilities in data generation and restoration, enabling it to precisely capture latent representations and effectively eliminate introduced noise.

As depicted in Figure 3 (bottom half), the Condition Latent Diffusion Model consists of a Stable Diffusion module introduced by (Rombach et al. 2022) and a Control Net branch from (Zhang, Rao, and Agrawala 2023). Stable Dif-

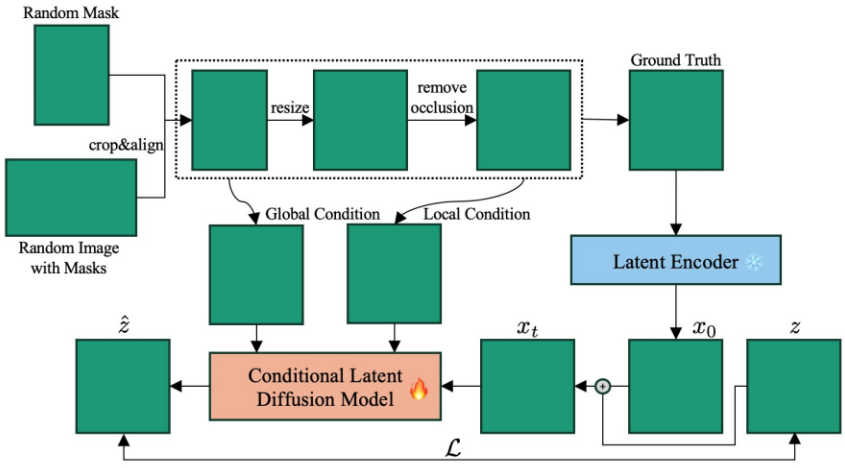


Figure 3: The pipeline of our proposed De-occlusion Method.

Algorithm 1: Training Procedure

Input: Original Image
Parameter: iteration
Output: Predicted noise z

masks = CropFormer(OriginalImage)
ConditionGlobal = masked(masks, OriginalImage)
Remove Complete Occlusion
ConditionGlobal = crop(ConditionGlobal)
Let iteration = iteration
while iteration **do**
    predictedZ = ControleNet(noiseZ, OriginalImage, ConditionGlobal, ConditionGlobal)
    loss(predictedZ, noiseZ)
    iteration = iteration - 1
end while

fusion module is applied for inpainting, aimed at restoring occluded regions, while the ControlNet branch is utilized to specify the exact regions to be restored, enhancing the overall inpainting efficacy.

Inference

As described in Algorithm 2, or a new input image, the first step involves using CropFormer to summarize the masks of all instances in the input image. Then, the process enters a loop: for each instance, the corresponding mask is removed from the image, and the region is repaired using an inpainter (Stable Diffusion with ControlNet). Subsequently, the repaired image is blended with the original image, and CropFormer is called to segment the blended image. The relationship between the generated masks (n2) at this point and the original masks (n1) is established by calculating the cost of their match.

Next, these generated masks are compared with the original masks. If the added area exceeds a predefined threshold, the original mask is replaced, effectively filling in the missing region.

Algorithm 2: Inference Procedure

Input: Original Image
Output: Recovered images

masks = CropFormer(OriginalImage)
Let iteration = masks.shape[0], i = 0
while iteration **do**
    recover(OriginalImage, masks[i])
    iteration = iteration - 1
end while

The core objective of this process is to iteratively repair instance regions, ensuring that the newly generated masks better match the original masks, thereby achieving precise image restoration.

Experiments

We now evaluate our method in various applications including ordering recovery, amodal completion, and amodal instance segmentation.

Datasets. 1) COCOA (Yan Zhu and Dollar 2017) is a subset of COCO2014 (Tsung-Yi Lin and Zitnick 2014) while annotated with pairwise ordering, modal, and amodal masks. We train PCNet on the training split (2,500 images, 22,163 instances) using modal annotations and test on the validation split (1,323 images, 12,753 instances). The categories of instances are unavailable for this dataset. Hence, we set the category ID constantly as 1 in training PCNet for this dataset. **2) EntitySeg**, originated from EntitySeg (Qi et al. 2023), it contains 33,227 images with high quality mask annotations. EntitySeg has three distinctive features. Firstly, 71.25% and 86.23% of the images are of high quality. Secondly, this dataset is open-world and not restricted to predefined classes. Thirdly, the mask annotation along the boundaries is more precise than those in other datasets.

Table 1: Comparison of evaluation metrics for PCNet trained on different datasets. 'acc_allp', 'acc_occp' and 'pAcc' refer to the accuracy of restoration without occlusion relationships, the accuracy of using occlusion relationships for restoration, and the accuracy of restoring all entities.‡indicates PCNet reproduced results

Train Dataset	**acc_allp**	**acc_occp**	**mIoU**	**pAcc**
PCNet-M	0.9601	0.8711	0.8134	0.8774
COCO‡	0.9576	0.8597	0.8090	0.8743
EntitySeg	0.9680	0.8721	0.8139	0.8859

Table 2: Amodal completion on COCOA validation, using ground truth modal masks.

method	amodal (train)	COCOA %mIoU
Supervised	✔	82.53
Raw	✘	65.47
Convex^R	✘	74.43
PCNet(NOG)	✘	76.91
PCNet(OG)	✘	81.35
Our	✘	81.60

Comparison Results

Dataset validity. To verify the high quality of the EntitySeg dataset, we only replaced the original PCNet dataset for experimentation, replacing the training set from coco2014 to cocoa_train, the testing will still be conducted on cocoa. In Table 1, we used pre-trained PCNet-M (Zhang, Rao, and Agrawala 2023) and compared the PCNet we replicated with the PCNet replicated on EntitySeg (Qi et al. 2023). The PCNet trained on EntitySeg showed a comprehensive improvement over the cocoa2014 metric, demonstrating the superiority of the EntitySeg dataset and its adaptability to occlusion removal tasks. Compared with COCO2014, which has a more similar distribution, the PCNet trained on EntitySeg also achieved better performance. **Amodal Completion.** We first introduce the baselines. For the supervised method, amodal annotation is available. A UNet is trained to predict amodal masks from modal masks end-to-end. Raw means no completion is performed. Convex represents computing the convex hull of the modal mask as the amodal mask. PCNet improves this baseline by using predicted order to refine the convex hull, constituting a stronger baseline: Convex^R. PCNet(NOG) represents the non-ordering-grounded amodal completion that relies on PCNet-M and regards all neighboring objects as the eraser rather than using occlusion ordering to search the ancestors. PCNet(OG) is PCNet ordering grounded amodal completion method.

We evaluate amodal completion on ground truth modal masks, as shown in Table 2. Our method surpasses the baseline approaches and PCNet method, which is comparable to the supervised counterpart.

Table 3: Ablation study of constructing data methods was conducted, and PCNet-M was trained on coco2014 and EntitySeg, respectively, to verify the effectiveness of our data construction method.

Train Dataset	modified	mIoU	pAcc
COCO 2014	✘	0.8090	0.8743
	✔	0.8132	0.8854
EntitySeg	✘	0.8139	0.8859
	✔	0.8197	0.8937

Ablation Study

In Table 3, the effectiveness of our data construction method was verified by training PCNet-m in two ways on COCO2014 and EntitySeg. Ticking the box indicates the use of our data construction method, and the results show that applying our method to different datasets can achieve better results.

Conclusion

To summarize, we present a novel paradigm for constructing self-supervised occlusion removal data. For the first time, we introduce the use of a diffusion model framework to address the occlusion removal task, aiming to approach the problem in a manner more aligned with human visual perception. We validate the feasibility of this method on open-world datasets, achieving excellent performance on the corresponding benchmarks.

However, the approach comes with increased computational costs, yielding only marginal improvements in performance. We also identify potential issues, such as the emergence of data distributions that do not adhere to physical laws, which are not present in the training set. In response, we are exploring an alternative approach by leveraging single-view 3D reconstruction to tackle these challenges comprehensively.

References

Avrahami, O.; Hayes, T.; Gafni, O.; Gupta, S.; Taigman, Y.; Parikh, D.; Lischinski, D.; Fried, O.; and Yin, X. 2022. SpaText: Spatio-Textual Representation for Controllable Image Generation.

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended Diffusion for Text-driven Editing of Natural Images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Brooks, T.; Holynski, A.; and Efros, A. 2022. InstructPix2Pix: Learning to Follow Image Editing Instructions.

Ehsani, K.; Mottaghi, R.; and Farhadi, A. 2018. SeGAN: Segmenting and Generating the Invisible. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2010. The Pascal Visual Object Classes

(VOC) Challenge. *International Journal of Computer Vision*, 303–338.

Gafni, O.; Polyak, A.; Ashual, O.; Sheynin, S.; Parikh, D.; and Taigman, Y. ???? Make-A-Scene: Scene-Based Text-to-Image Generation with Human Priors.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*.

Gong, K.; Liang, X.; Zhang, D.; Shen, X.; and Lin, L. 2017. Look into Person: Self-supervised Structure-sensitive Learning and A New Benchmark for Human Parsing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kar, A.; Tulsiani, S.; Carreira, J.; and Malik, J. 2015. Amodal Completion and Size Constancy in Natural Scenes. *Cornell University - arXiv,Cornell University - arXiv*.

Kimia, B. B.; Frankel, I.; and Popescu, A.-M. 2003. Euler spiral for shape completion. *International Journal of Computer Vision*, 54(1-3): 159–182.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; Duerig, T.; and Ferrari, V. 2020. The Open Images Dataset V4. *International Journal of Computer Vision*, 1956–1981.

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, K.; and Malik, J. Springer,2016. Amodal instance segmentation. *European Conference on Computer Vision, pages 677–693*, 1,2.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. 2023. GLIGEN: Open-Set Grounded Text-to-Image Generation.

Lin, H.; Wang, Z.; Feng, P.; Lu, X.; and Yu, J. 2016. A computational model of topological and geometric recovery for visual curve completion. *Computational Visual Media*, 2(4): 329–342.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. *Microsoft COCO: Common Objects in Context*, 740–755.

Lu Qi, S. L. X. S., Li Jiang; and Jia, J. June 2019. Amodal instance segmentation with kins dataset. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1, 2, 5.

Meng, C.; He, Y.-L.; Yang, S.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. *Cornell University - arXiv,Cornell University - arXiv*.

Nanay, B. 2018. The importance of amodal completion in everyday perception. i-Perception. *Iperception*, 9(4): 2041669518788887.

Omer, B.-T.; Lior, Y.; Yaron, L.; and Tali, D. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. *Cornell University - arXiv,Cornell University - arXiv*.

Patrick Follmann, P. H. a. R. M. K., Rebecca Ko Nig; and Ttger, T. B. IEEE, 2019. Learning to see ¨ the invisible: End-to-end trainable amodal instance segmentation. *2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1328–1336*, 1,2.

Qi, L.; Jiang, L.; Liu, S.; Shen, X.; and Jia, J. 2019. Amodal Instance Segmentation With KINS Dataset. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qi, L.; Kuen, J.; Guo, W.; Shen, T.; Gu, J.; Jia, J.; Lin, Z.; and Yang, M.-H. 2023. High-Quality Entity Segmentation. arXiv:2211.05776.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. *Playing for Data: Ground Truth from Computer Games*, 102–118.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2022. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation.

Silberman, N.; Shapira, L.; Gal, R.; and Kohli, P. 2014. A contour completion model for augmenting surface reconstructions. In *European Conference on Computer Vision (ECCV)*.

Tsung-Yi Lin, S. B. J. H. P. P. D. R. P. D., Michael Maire; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. *ECCV*, 1, 2, 3, 4, 5.

Voynov, A.; Aberman, K.; and Cohen-Or, D. 2022. Sketch-Guided Text-to-Image Diffusion Models.

Wrenninge, M.; and Unger, J. 2018. Synscapes: A Photorealistic Synthetic Dataset for Street Scene Parsing. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*.

Xiaohang Zhan, B. D. Z. L. D. L., Xingang Pan; and Loy, C. C. June 2020. Self-supervised scene deocclusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1, 2, 5.

Yan Zhu, D. M., Yuandong Tian; and Dollar, P. 2017. Semantic amodal segmentation. *CVPR*, 1464–1472.

Zhan, X.; Pan, X.; Dai, B.; Liu, Z.; Lin, D.; and Loy, C. C. 2020. Self-Supervised Scene De-Occlusion. 3783–3791.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543.

Zhao, J.; Li, J.; Cheng, Y.; Sim, T.; Yan, S.; and Feng, J. 2018. Understanding Humans in Crowded Scenes: Deep Nested Adversarial Learning and A New Benchmark for Multi-Human Parsing. In *Proceedings of the 26th ACM international conference on Multimedia*.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene Parsing through ADE20K Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhu, Y.; Tian, Y.; Metaxas, D.; and Dollar, P. 2017. Semantic Amodal Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.