Open-World Amodal Appearance Completion

Supplementary Material

This supplementary material provides additional details, analyses, and resources to complement the main paper. Specifically, we include additional visual comparisons, showcasing more examples that compare our method with existing approaches in Sec. 1. We then provide a failure analysis in Sec. 2, categorizing failures into two types: complete failures, where no amodal completion is generated, with proportions of such failures for our method and comparison methods; and unsatisfactory completions, where the generated amodal result deviates from the expected appearance. In Sec. 3, the human study details are expanded upon with sample questionnaires used in the evaluations and inter-annotator agreement metrics, such as Fleiss' kappa, to establish the reliability of subjective assessments. The dataset collection process is described in detail in Sec. 4, outlining the steps taken to create our evaluation dataset and including visualizations to demonstrate dataset diversity. Lastly, comprehensive configuration details are provided in Sec. 5, including the specifications of pre-trained models used, to support reproducibility and facilitate further research.

1. Additional Visual Comparisons

The visual comparisons in Fig. 3 showcase the effectiveness of our method in handling diverse and complex scenarios across a wide range of object categories and occlusion types. The examples span indoor, outdoor, and natural scenes, demonstrating our approach's open-world adaptability and robustness in producing realistic and complete amodal completions.

Our approach excels in reconstructing objects from diverse categories, including buildings, furniture, animals, insects, and tools. This adaptability underscores its ability to handle open-world scenarios without predefined object categories, unlike competing methods such as PD w/o MC and PD-MC [12], which depend on fixed classes and fail when encountering unseen categories. In contrast, Pix2gestalt [8], while always producing amodal completions due to its reliance on supervised learning with large training datasets, sometimes minimally alters the input image (e.g., the moth and cat examples), offering little meaningful reconstruction.

Our method demonstrates the ability to handle complex occlusions, such as objects occluded by shrub or fences, mutual overlapping elements, or ambiguous background regions. Examples such as the building, moth, and cat emphasize the critical role of our background segmentation strategy in identifying and effectively handling occluding background segments. This strategy improves structural integrity in scenarios where competing methods fail entirely (e.g., PD w/o MC and PD-MC) or produce incomplete reconstructions (e.g., Pix2gestalt).

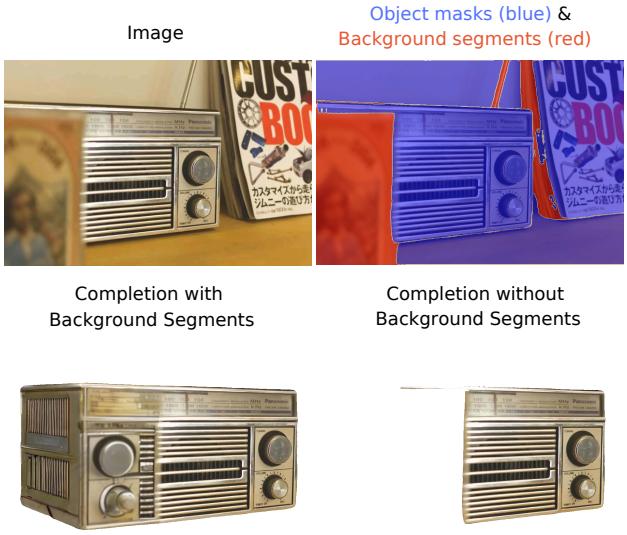

Figure 2. Comparison of our amodal completion results with and without considering background segments. Object masks are shown in blue, and background segments are highlighted in red. Our method with background segmentation reconstructs the occluded object more comprehensively, capturing missing structural details (e.g., the side panel of the radio) that are ignored when background segments are not utilized.

Impact of Background Segmentation. Fig. 2 demonstrates the importance of incorporating background segments into our method. Our background segmentation strategy successfully identifies and accounts for ambiguous background regions, such as the red-highlighted areas. With background segments considered, the amodal completion preserves structural integrity and reconstructs missing details, such as the side panel of the radio.

2. Failure Analysis

We analyse failure cases in two types: (i) complete failures, where no amodal completion is generated, and (ii) partial failures, where the generated completion is unsatisfactory.

Complete Failures. Tab. 1 details the proportion of complete failures—cases where no amodal completion result is generated—for each method across datasets. While Pix2gestalt achieves a 0% failure rate due to its supervised design, it sometimes minimally alters the input without meaningfully addressing occlusion, as seen in Fig. 3 (e.g., moth and cat examples). In contrast, our method maintains

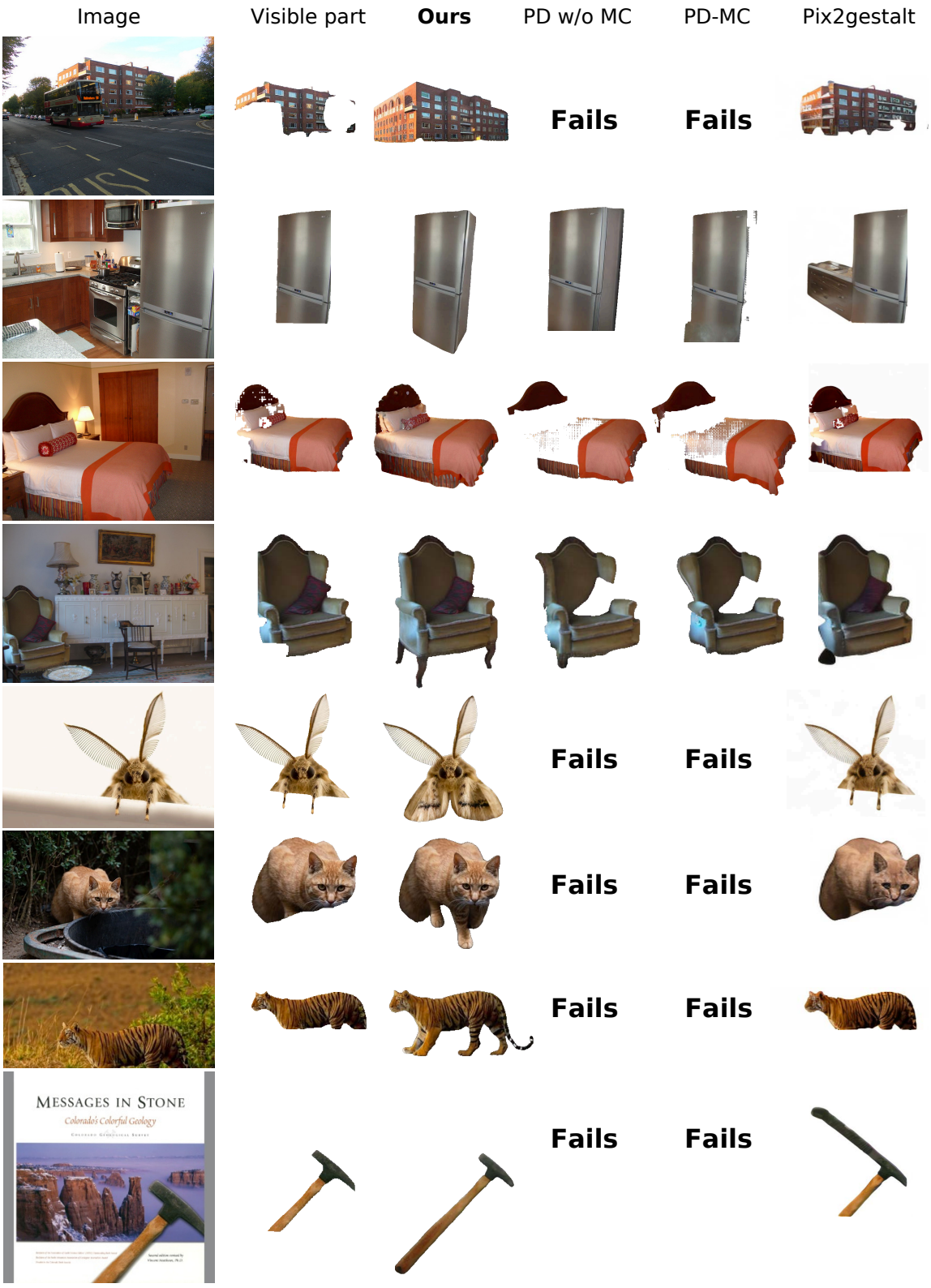

Figure 3. Visual comparisons of amodal completions across different methods. Our method consistently outperforms others in reconstruction integrity, handling complex occlusions, and producing plausible completions. Examples are drawn from VG, COCO-A, free images, and LAION datasets, with two images from each source. "Fails" refers in cases where no amodal completion result is generated.

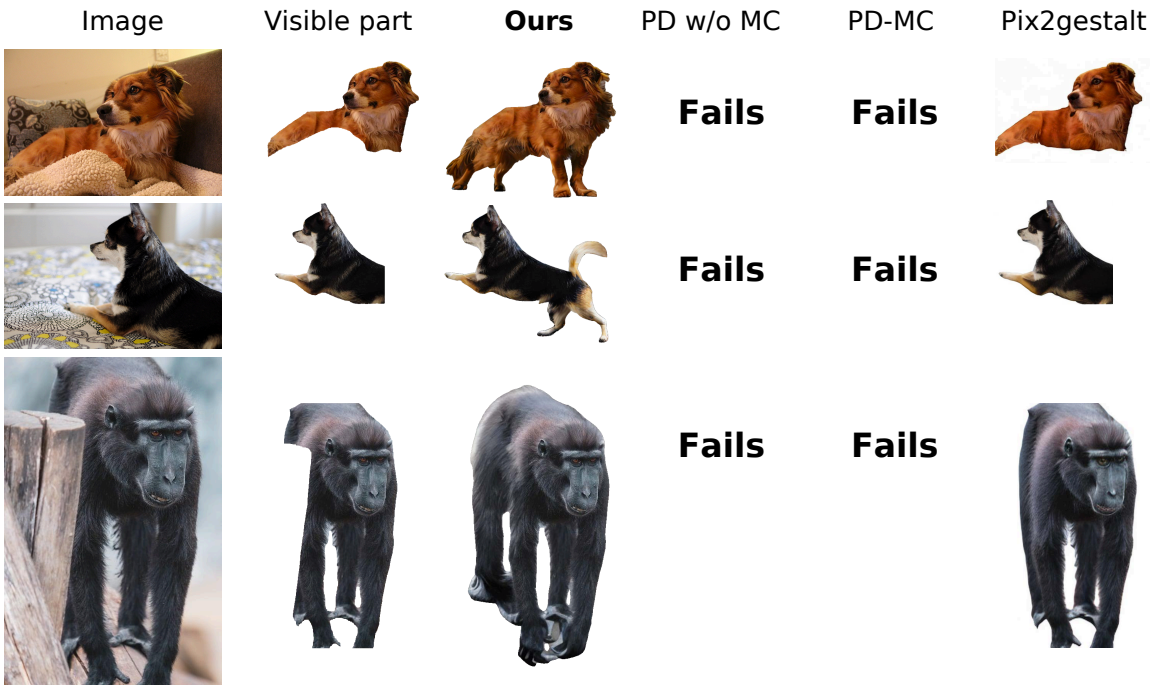

Figure 4. Examples of our common partial failures, such as unnatural poses and hand reconstruction. "Fails" refers in cases where no amodal completion result is generated. Other methods have either completely failed or partially failed in these cases. Compared to other methods, our framework maintains structural integrity, even if some results are not satisfactory.

	Training Free	VG [3]	COCO-A [13]	Free Image	LAION [11]	Overall
PD w/o MC	Yes	39.6%	47.3%	58.2%	53.1%	44.9%
PD-MC	Yes	40.0%	48.5%	58.2%	54.2%	45.5%
Pix2gestalt	No	0.0%	0.0%	0.0%	0.0%	0.0%
Ours	Yes	6.3%	1.8%	2.0%	1.7%	4.1%

Table 1. Proportion of complete failures (no amodal completion generated) for each method across datasets.

a low failure rate (4.1% overall) while ensuring structural integrity and realistic completions. Unlike PD w/o MC and PD-MC, which exhibit overall failure rates exceeding 40% across datasets, our approach succeeds in reconstructing occluded objects even under challenging conditions. Our failures occur primarily in cases where occluded objects are undetectable. Notably, our method avoids meaningless pixel modifications by not attempting amodal completions when occluded objects are undetectable.

Partial Failures. While our method excels in most cases, occasionally partial failures occur, where the generated completion does not match expectations. Fig. 4 showcases common partial failures, such as generating a standing dog instead of a sitting one based on the context, or producing gorilla hands with unnatural shapes. These issues arise primarily due to limitations in pre-trained models used for inpainting and the inherent ambiguity in certain occlusions. Unlike other methods, however, our framework maintains structural integrity, even in partially unsatisfactory results. Future work could address these limitations by refining inpainting models and incorporating additional contextual reasoning mechanisms.

3. Human Study Design and Inter-Participant Agreement

To evaluate the subjective quality of amodal completions, we conducted a structured human study[1] designed to assess the realism and completeness of generated results across various methods. Using the Prolific crowdsourcing platform[2], we recruited 180 participants to compare our method against Pix2gestalt [8], PD w/o MC, and PD-MC [12]. Each participant was presented with an original image alongside four completed versions corresponding to the methods under evaluation. The order of the four completions was randomized to mitigate positional bias. Participants were tasked with selecting the completion that best represented a realistic and whole view of the object, based on visible cues from the original image.

To ensure the task's clarity, we provided participants with a detailed guide outlining the evaluation process

[1]This study has received Human Ethics Approval (2024-30689-58436-3) from the University of Melbourne.

[2]https://www.prolific.com/

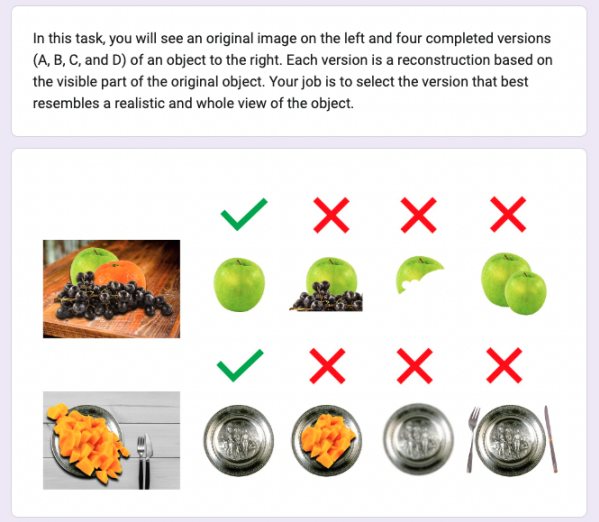

Figure 5. A detailed guide provided to participants at the beginning of the questionnaire, demonstrating how to assess the realism and completeness of amodal completions

(see Fig. 5). Additionally, Fig. 6 illustrates a sample question interface. To ensure data reliability, 10% of the questions were "gold standard" trials with unambiguous correct answers. Only participants who passed at least 75% of these quality control checks were included in the final analysis. The "gold standard" trials were not included in the main data analysis.

Inter-Participant Agreement. To quantify the inter-participant agreement across multiple raters, we computed Fleiss' kappa scores for each dataset and the overall study. Tab. 2 summarizes the calculated κ values, which consistently fall within the "fair agreement" range [5]. Among the datasets, the highest agreement was observed for the LAION (κ = 0.374), while the overall agreement for the study was κ = 0.319. This indicates a fair level of consistency among participants despite the inherent subjectivity of the task.

VG [3]	COCO-A [13]	Free Image	LAION [11]	Overall
0.275	0.364	0.353	0.374	0.319

Table 2. Fleiss' kappa between human participants.

The observed agreement reflects the complexities involved in assessing amodal appearance completions, which require judgments on both perceptual realism and context-dependent plausibility. While variability in individual preferences is expected, the fair consistency across all datasets highlights the reliability of our evaluation framework. This findings also shows the importance of subjective evaluation in capturing perceptual nuances that quantitative metrics may overlook.

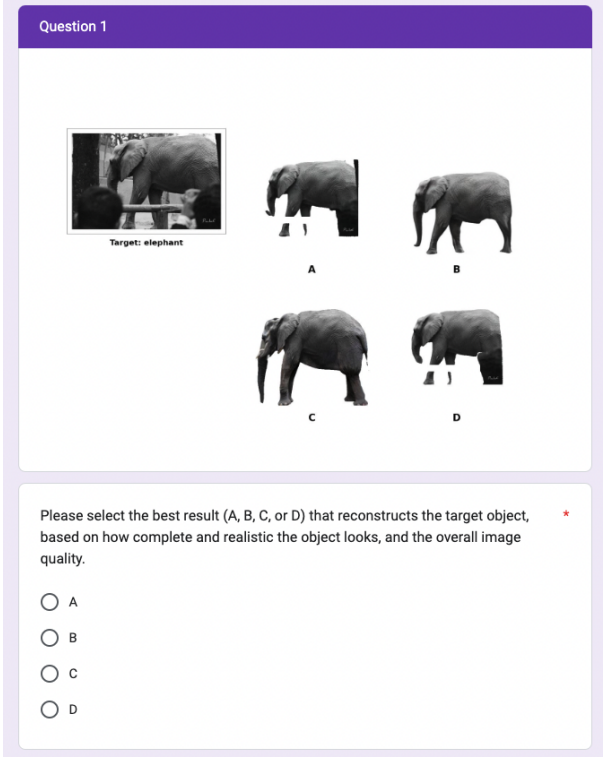

Figure 6. An example question from the human study, showing the original image (left) and four completions in randomized order (right). Participants were instructed to select the version that best reconstructed the target object.

4. Dataset Collection

Our evaluation dataset integrates images from four sources: COCO-A [13], Visual Genome (VG) [3], LAION [11], and copyright-free images collected from publicly accessible websites[3][4][5]. A total of three human annotators collected images containing occluded objects from different sources, and each independently provided a category label for one or more occluded objects in the image. The resulting dataset consists of 2379 images spanning 553 distinct target object classes.

COCO-A contribute natural scenes with realistic occlusions, providing a foundation of everyday scenarios and common objects. However, not all images in COCO-A feature object-specific occlusions since it was originally designed for semantic segmentation [13]. To address this, we applied a filtering process that removed images where (a) background elements were occluded but primary objects were not, (b) the visible part of the primary object occupied less than 2% of the total image area, (c) most of the primary object lay outside the image boundary, or (d) occluders were

[3] https://www.pexels.com/
[4] https://pixabay.com/
[5] https://unsplash.com/

Figure 7. Examples of filtered images from COCO-A [13]: (a) Background elements occluded but primary objects visible (e.g., zebra, airplane). (b) Visible object area below 2% of the total image area (e.g., surfboard, vehicle). (c) Most of the primary object lies outside the image boundary. (d) Occluders are transparent or excessively thin (e.g., glass, wires).

transparent or excessively thin (e.g., glass or wires). Fig. 7 illustrates examples of filtered images from COCO-A.

To further enhance diversity, we incorporated images from VG [3], LAION [11] and copyright-free sources, introducing a broad range of lighting conditions, object appearances, and complex occlusions typical of unconstrained, open-world environments. Fig. 8 shows the distribution of images across the four dataset sources. VG accounts for the largest share at 51.9%, followed by COCO-A (31.6%), copyright-free images (9.6%), and LAION (7.0%). The combination of these sources ensures that our evaluation dataset captures a wide range of real-world occlusion scenarios and object categories.

To analyse our evaluation dataset composition, we further grouped object labels into broad categories based on their semantic meanings. This grouping provides insights

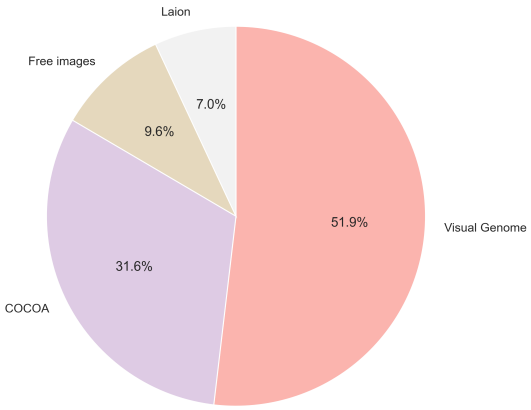

Figure 8. Image distribution across different datasets.

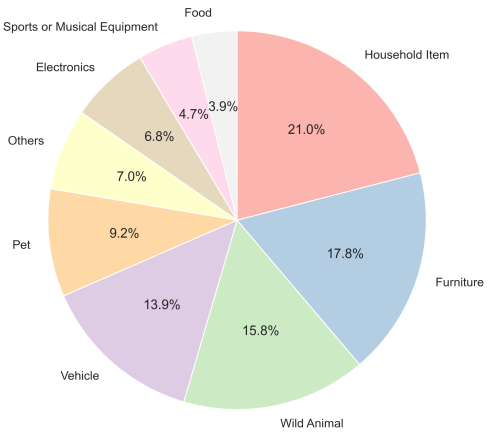

Figure 9. Distribution of object categories in major groups.

into the types of objects present in the dataset. As shown in Fig. 9, prominent categories include "Household Item" (21%), "Furniture" (17.8%), and "Wild Animal" (15.8%), reflecting the dataset's relevance to both everyday scenarios and naturalistic environments. Other categories, such as "Vehicle", "Pet", and "Food" further ensure coverage of real-world contexts across various settings.

The grouping of object labels into these broad categories was based on specific mapping rules. For instance, "Household Item" encompasses items commonly found in daily home life, including objects like bowls, spoon and scissors. The "Furniture" category includes various types of household furnishings such as sofas, chairs and tables. Domesticated animals such as dogs and cats were categorized under "Pet", while "Wild Animal" includes non-domesticated species like elephants, giraffes and bear. Transportation-related objects, such as trucks, bicycles, and airplanes, were grouped into the "Vehicle" category, whereas electronic devices like laptops, phones and cameras were placed in "Electronics". The "Sports or Musical Equipment" cate-

gory includes items like tennis rackets, guitars, and piano, covering recreational and artistic tools. "Food" represents a variety of edible items, including apples, bread, and pizza. Finally, a catch-all "Others" category includes objects that do not fit neatly into the previous groups, such as buildings and natural elements (e.g., trees, flowers).

5. Configuration Details

To ensure reproducibility and facilitate further research, this section details the configuration and pre-trained models used in our framework.

For vision language-grounded object identification, we utilized LISA-13B-llama2-v1 model [4], which offers robust reasoning capabilities for mapping natural language queries to visible object regions. The appearance of the occluded regions were reconstructed using the Stable Diffusion v2 inpainting model [10], known for its high-quality generative performance. To enhance scene understanding and support object detection, we incorporated the RAM++ image tagging model (ram_plus_swin_large_14m) [1], enabling open-set tagging of visual elements, and the GroundingDINO object detector (groundingdino_swint_ogc) [7], which effectively identifies and segments objects in open-world settings.

To assess occlusion relationships, we employed InstaOrderNet (InstaOrder_InstaOrderNet_od) [6], a model pre-trained for amodal occlusion ordering tasks. This model processes pairwise object masks and image patches without relying on object category labels, making it suitable for the diverse and ambiguous occlusions in open world scenes. For pixel-wise segmentation tasks, we used the Segment Anything model (sam_vit_h_4b8939) [2], which provided accurate segmentation across various object and background types. The CLIP model (ViT-B/32) [9] was utilized for text-image alignment during inpainting prompt generation, leveraging its powerful feature extraction for matching visual and semantic cues.

These pre-trained models, each specializing in a specific subtask, enabled us to construct a robust framework tailored to the challenges of open-world amodal appearance completion. Leveraging their embedded knowledge allowed us to address complex, real-world scenarios without requiring additional training. Furthermore, the modularity of our framework ensures that each component can be easily replaced with improved pre-trained models as they become available, enhancing adaptability and future extensibility.

References

[1] Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. Open-set image tagging with multi-grained text supervision. *arXiv e-prints*, pages arXiv–2310, 2023. 6

[2] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 6

[3] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2017. 3, 4, 5

[4] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 6

[5] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977. 4

[6] 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. 6

[7] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*. Springer, 2024. 6

[8] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3931–3940. IEEE Computer Society, 2024. 1, 3

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 6

[10] 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. 6

[11] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 3, 4, 5

[12] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Amodal completion via progressive mixed context diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9099–9109, 2024. 1, 3

[13] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1464–1472, 2017. 3, 4, 5