An Empty Room is All We Want:
Automatic Defurnishing of Indoor Panoramas

Mira Slavcheva⋆   Dave Gausebeck⋆   Kevin Chen⋆   David Buchhofer   Azwad Sabik   Chen Ma
Sachal Dhillon   Olaf Brandt   Alan Dolhasz

Matterport

Abstract

We propose a pipeline that leverages Stable Diffusion to improve inpainting results in the context of defurnishing—the removal of furniture items from indoor panorama images. Specifically, we illustrate how increased context, domain-specific model fine-tuning, and improved image blending can produce high-fidelity inpaints that are geometrically plausible without needing to rely on room layout estimation. We demonstrate qualitative and quantitative improvements over other furniture removal techniques.

1. Introduction

The recent advancement of digital technologies has revolutionized various industries, and real estate is no exception [40]. The emergence of *digital twins*—virtual replicas of physical assets—has significantly impacted the way properties are marketed, managed, and visualized [16]. Digital twins offer a plethora of benefits, including enhanced decision-making [24], improved design processes [9], and immersive experiences for stakeholders [12].

While digital twins excel in replicating physical structures with remarkable accuracy, they often lack the flexibility to adapt to different downstream scenarios and requirements. One critical aspect that remains underexplored is *defurnishing*: removing furnishings and objects from virtual representations of built environments [14]. Defurnishing offers numerous advantages, such as enabling potential buyers or renters to envision personalized layouts [47], facilitating interior design experimentation [22], and providing insights for property evaluation and renovation [7]. It is a fundamental component of any workflow dealing with digital twins, particularly when captured from a built environment rather than synthesized from the ground up.

In this paper, we focus on defurnishing indoor panorama images, leaving the study of defurnishing other common

---

⋆ denotes equal contribution.
✉ research@matterport.com

Figure 1. **Indoor panorama defurnishing.** Our custom fine-tuning of Stable Diffusion inpainting reduces its tendency to hallucinate objects near shadows and reflections, such as the radiators on the walls and lamp in the corner.

representations of digital twins, such as polygon meshes, to future work. Additionally, while a digital twin may be composed of several posed panorama images, we focus on defurnishing single images in isolation and do not consider consistency between multiple views.

Recent advances in generative models, such as latent diffusion models [36], have dramatically improved the quality of image inpainting [10] by leveraging learned contextual priors to better complete the missing information in an image. However, as these models are commonly trained to inpaint objects and scene content rather than empty space, spurious inpaints or *hallucinations* [28, 59] tend to be a significant problem when using these models in the context of defurnishing. Specifically, Stable Diffusion attempts to

"explain away" contextual information such as shadows and object parts left behind from small imperfections in the inpainting mask by placing objects where empty space was expected (see Figure 1). This issue can be partially resolved by applying morphological operations such as dilation to the inpainting mask [20], but this does not adequately handle all cases of shadows, light beams, and other similar effects, which are related to the furniture but not co-located with it [53].

Consequently, we present a system for automatic furniture removal, at the core of which is a custom inpainting model that builds over Stable Diffusion in the following ways:

1. Leverages a custom dataset of unfurnished panoramas and an augmentation strategy to make the model robust to imperfect masks, cast light beams, reflections, and shadows, significantly reducing hallucinations.
2. Inpaints equirectangular images natively, significantly increasing the amount of available contextual information compared to perspective images.
3. Incorporates superresolution and a custom blending strategy to achieve a high-quality, seamless result.

2. Related work

The key component of our method is an improved inpainting routine, so here we focus on works related to this task.

2.1. Inpainting

Digital inpainting is the task of generating pixels to complete missing regions of an image [2]. Many inpainting methods have been developed over the past decades, ranging from approaches using classical techniques [1, 2, 6, 17, 33, 44] to those leveraging deep neural networks [19, 27, 32, 34, 36, 41, 43, 50–52].

Early learned approaches using Convolutional Neural Networks (CNNs) pioneered the use of Generative Adversarial Networks (GANs) for inpainting [19, 34, 50]. Partial convolutions [27] and their learned counterparts, gated convolutions [51], allowed inpainting with arbitrary mask shapes by propagating masks through the network layers. Many models employed a two-stage pipeline that generated intermediate representations to aid in the final inpainting step, such as coarse resolution inpaints [50–52], edge maps [32], and semantic segmentation masks [41]. To handle large masks, fast Fourier convolutions [43] were shown to increase the effective receptive field of the neural network and aid in understanding the global structure of the image.

2.2. Panoramic furniture removal

Given a panoramic image of an indoor scene, furniture removal generally involves two important steps [14]:

1. Segment all instances of furniture to remove.

2. Inpaint on the combined furniture masks.

While an off-the-shelf semantic segmentation model [4, 11, 29, 37, 57] trained on a suitable ontology [58] is sufficient for the first step, a number of different approaches [13–15, 21, 25] have been explored to obtain realistic inpaints using CNN GAN inpainting models. PanoDR [14] proposed that in order to produce inpaints that preserve real-life structures, predicting the room layout (*e.g.* segmentation for floor, walls, and ceiling) is a crucial pre-processing step. To address perspective distortion and mixing textures coming from different surfaces, Kulshreshtha *et al.* [25] used the room layout to separate the room into distinct 3D planes [26], then inpainting each plane separately after orthographic reprojection. Gao *et al.* [13] improved upon this by performing inpainting on each 3D plane in a single end-to-end pipeline.

These approaches all leverage a room layout estimator [42, 56] to explicitly or implicitly guide inpainting. This makes it difficult to generalize these models to handle certain features common to real-life room layouts, such as curved walls, kitchen islands and half walls, and staircases. With recent advances in image inpainting techniques, we study whether structural cues from room layout estimation are necessary for creating high-fidelity and geometrically plausible inpaints.

2.3. Stable Diffusion

Latent diffusion models, such as Stable Diffusion (SD) [36], have recently risen to the forefront of image generation and demonstrated results that out-perform their GAN counterparts [10]. Diffusion models are readily scalable to better model the complex distribution of training data and can sample diverse inpaints at high fidelity [30]. SD is a text-to-image model trained on a large image dataset [39] that can also be conditioned by multi-modal inputs, including line contours, depth maps, and other images for image-to-image translation [54].

Similar to other generative AI models, like large language models, SD suffers from *hallucinations*—generation of plausible but incorrect information—due to issues such as training data impurities and bias, and attention dilution at inference [45, 49]. Understanding why hallucinations occur and how they can be mitigated are research areas that are critically important for using these generative models for downstream tasks and in our daily lives.

In the case of inpainting for defurnishing, one root cause for hallucinations, which we hypothesize is very significant, is data bias and object co-occurrence [28, 59]. If the dataset used to train SD contains more images of furnished indoor rooms than unfurnished ones, then when the trained model is presented with an image of an indoor room, it would be biased towards generating inpaints including furniture. In other words, the dataset may contain a bias where an
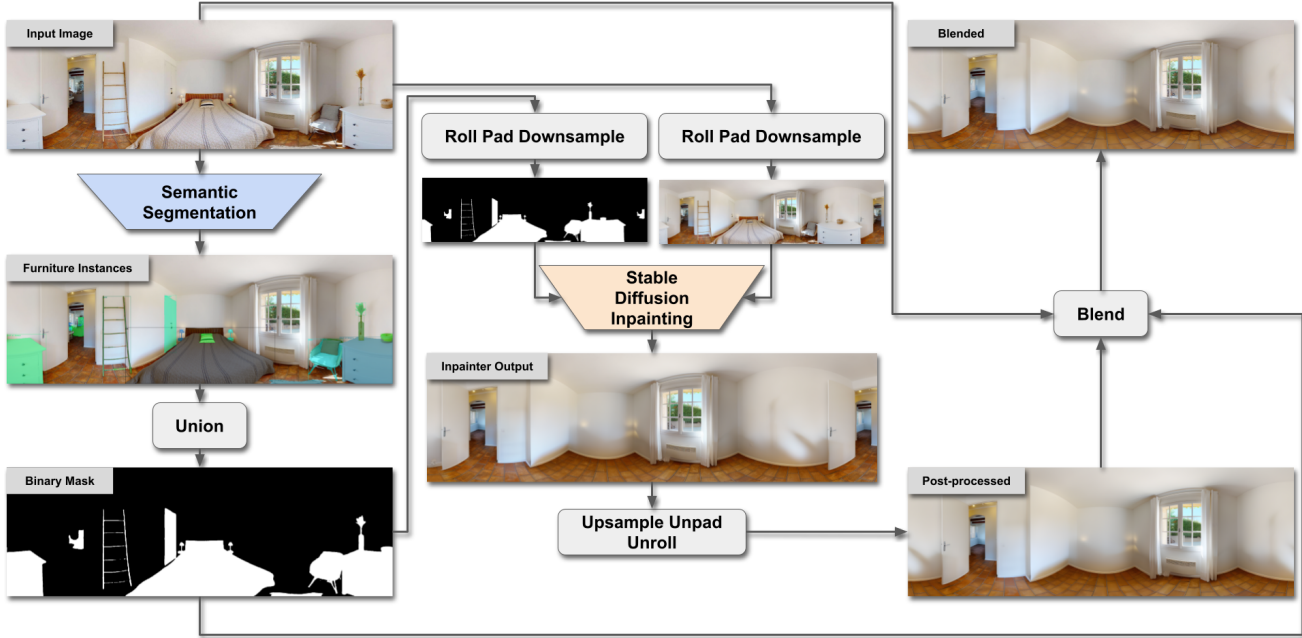
Figure 2. **Defurnishing pipeline.** The input to our system is a 8192×4096 pixel equirectangular panorama, which we crop to 3:1 aspect ratio to exclude the poles. **Pre-processing:** We obtain a binary furniture mask via semantic segmentation. Both input and mask are rolled, so that inpainting regions are in the center of the image, and padded, to ensure sufficient context (note the repeated doorway and cupboard in the example). The images are then downsampled to a height of 512 pixels. **Inpainting:** Our custom process is robust to inexact masks and remnant shadows, as detailed in the method section. **Post-processing:** We apply 4× superresolution to the inpainted output and then blend the original and inpainted panoramas into the final result using the mask, keeping as much of the original resolution details as possible.

"indoor room" and "furniture" often co-occur leading the model to learn this association and spuriously hallucinate furniture in images of indoor rooms.

Regarding dataset bias, one remedy would be to choose text prompts that guide the inpaint towards the desired outcome. Another approach is to fine-tune [18, 38] the diffusion model on another dataset where this bias has been rectified, trading diversity for fidelity in a particular use-case.

3. Method

Our method takes a high resolution panorama image as input and outputs a defurnished version of it, as visualized in Figure 2. We start by semantic segmentation of all furniture. Next, we run our custom unfurnished space inpainting. As it is SD-based, it is done at a lower resolution, so finally we run superresolution and blend the inpainted and original images in order to obtain a visually pleasing final result. We elaborate on each of these steps in the following.

3.1. Pre-processing

Furniture segmentation We use an off-the-shelf semantic segmentation network [4, 23] to identify all instances of furniture classes, as defined in the ADE20K benchmark [58]. Their union is our inpainting mask. Note that, in our experience, modern semantic segmentation networks

work equally well on equirectangular and perspective images. This spares us the back-and-forth conversion between a panorama and a set of perspective images that other methods necessitate [21], yielding a continuous mask and faster processing in our case.

Context maximization Next, we ensure that our image is set up optimally for inpainting. To this end, we roll the panorama around horizontally so that the masked objects are as close to the center as possible. There is no guarantee that there will be no masked object on the image edge, so we additionally wrap-pad the image horizontally to provide sufficient context for the ensuing inpainting. We apply the same transformations to the corresponding mask image.

3.2. Robust unfurnished space inpainting

Figure 1 points to three major issues with SD inpainting that we aim to solve:
1. It is not readily applicable to non-square images.
2. It is prone to hallucinating objects.
3. It is low resolution and lacks high frequency details.
Hallucinations are a known issue for large generative models [28, 49]. The same prompt gives different results with different random seeds, some of which do not exhibit hallucinations, but this is unreliable in an automated system.

Figure 3. **Training dataset examples.** Synthetic furniture items and shadows are rendered over real unfurnished panoramas.

This issue is multi-faceted in the case of inpainting. First, inpainting masks may be inaccurate, leaving visible parts of objects which SD picks up and completes into full objects. While this can be addressed with mask dilation, valuable context and detail is lost. Even with perfect object masks, there are shadows left, which SD aims to explain by inpainting objects that may have cast them. Second, our case of inpainting empty space is especially challenging, because generative models are good at creating concrete objects with describable properties (such as round wooden kitchen tables), but struggle with generating imagery for vague concepts like *emptiness*. Negative prompts are another way, but while they can make hallucinated objects look less like the negative prompt, they do not remove the objects.

We fine-tune a version of SD inpainting to address the first two issues. For the third one, we develop a post-processing routine described in Section 3.3.

Dataset To be able to successfully inpaint *empty space*, we train on a proprietary dataset of 160k equirectangular panoramas of unfurnished residential spaces, similar to Matterport3D [3, 35]. As discussed, the use of equirectangular panoramas ensures that we maximize the context at inpainting time. We found that SD quickly adapts to this kind of imagery after being exposed to fewer than 1000 panoramas. We find this strategy to be better than applying SD to a tiled version of the equirect, because all context is present in one image.

Synthetic data & augmentations To make inpainting robust to shadows and inaccurate masks, we augment the empty spaces with synthetically rendered furniture objects from a dataset like Objaverse [8]. The rendering is not photorealistic, but includes shadows, as shown in Figure 3. We fine-tune a pre-trained SD inpainting model using as inputs the unfurnished space panoramas with synthetically rendered objects and shadows, and masks that only cover the objects and not the shadows. The target outputs are the original unfurnished panoramas. This fine-tuning discourages SD from hallucinating objects when seeking to explain away effects of the furniture on the scene, such as shadows, reflections, or light beams. Additionally, to make our model robust to mask inaccuracies, we perturb the input masks to simulate imperfect semantic segmentation output.

Figure 4. **Effect of number of training prompts.** Fewer than 32 prompts lead to hallucinations near the shadow on the right wall.
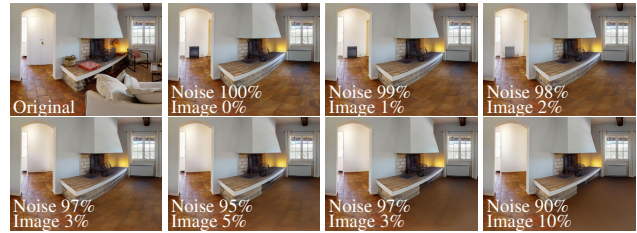

Figure 5. **Initialization of latents for inpainting.** With decreasing percentage of noise in the initialization, hallucinations in the hallway decrease, but blurriness on the inpainted floor increases.

Prompts SD is a text-to-image model, and consequently the text prompt(s) used for our inpainting task play(s) an important role. The prompt we use at inference time is *empty room*, but we found that training with a set of similar prompts further reduces hallucinations, as shown in Figure 4. We randomly select one prompt from the set per training sample (prompts for the same sample may differ in different epochs). The additional prompts contain synonyms, such as unfurnished; space, home, house; as well as appending descriptions such as "uniformly blank". We tested with 0 (empty prompt), 1, 8, 32, 72 and 180 prompts. In our experiments 0-8 prompts lead to models that are most prone to hallucination, 32 and 72 yield results with much fewer hallucinations, while more prompts result in increased hallucinations again. Thus, in our experiments we choose to train with 32 variants of the "empty room" prompt.

Initialization At inference time we initialize the inpainting latents as a weighted sum of 97% random noise and 3% latents based on the input image, blurred under the inpainting mask. We empirically found this combination to best minimize both hallucinations and blur, as shown in Figure 5.

Resources We train for 96 hours on 8 NVIDIA A10G GPUs (24 GB vRAM) with an effective batch size of 96 (using gradient accumulation). The base SD version that we fine-tune is *stabilityai/stable-diffusion-2-inpainting* and we only perform low-rank adaptation (LoRA) [18], whereby

we calculate update matrices for its UNet, rather than training all its weights. We experimented with training the entire UNet and the VAE, each of which gave inferior results. This could be due to our dataset size or other factors, which we did not investigate further. We also fine-tuned SD XL, but found it to be more prone to hallucinations.

We only use one GPU for inference. We found that 10 inference steps are sufficient. Our pipeline takes approximately 12 seconds to process a single panorama: 8s for pre-processing and semantic segmentation, 3.5s per image for inpainting alone, and 0.5s for post-processing.

3.3. Post-processing

After inpainting, we apply an off-the-shelf superresolution network [46] to upscale the inpainted image four times. We then undo the padding and rolling pre-processing transformations to restore the original panorama dimensions. Finally, we make sure that the inpainted and original textures match well via a custom blending routine.

Blending SD-based inpainting tends to lack high frequency details, which is problematic in our setting, where a high-resolution final output is required. Therefore, we aim to preserve as much detail as possible from the original furnished panorama, and develop a tailored blending procedure that uses the binary inpainting mask to select how to combine the original and inpainted images. As explained in Section 3.2, our inpainting strategy is guided by the inpainting mask, but is trained to be robust to shadows by allowing for pixels nearby the mask to also be modified, as enforced by the dataset and training objective. In this way, we not only avoid hallucinating objects due to the remaining shadow, but we also remove the shadow itself from the image, leading to an overall higher-quality inpainted result. Conversely, if we were to directly use the mask for blending the original and inpainted panoramas, shadows would be re-introduced in the final image. Therefore, we use the pixels from the inpainted version not only where the mask indicates, but also in nearby regions where the inpainted image is significantly different from the original. Similarly, if those significant changes are far away from the inpainting mask, they are more likely to be spurious hallucinations, so we use pixels from the original image. Figure 8 visualizes the benefits of this strategy.

4. Results

In this section, we compare to related works and analyze the contributions of the various components of our pipeline.

4.1. Furniture removal

We first analyze only the custom-trained inpainting component of our pipeline. The input consists of a 1536×512 pixel

Method	PSNR ↑	SSIM ↑	JOD ↑	LPIPS ↓
LaMa [43]	**26.76**	**0.916**	7.670	0.094
LGPN-Net [13]	25.57	0.867	7.439	0.206
SD-2-inpaint	24.67	0.878	7.487	0.101
Ours-inpaint	26.02	0.874	**7.691**	**0.091**
Ours-full	**27.05**	**0.930**	**7.753**	**0.056**

Table 1. **Quantitative evaluation.** Our inpainting module is slightly worse than LaMa [43] on absolute difference metrics, because the underlying SD outputs lower-frequency textures. Together with superresolution and blending, our full pipeline outperforms all related techniques.

panorama image and a corresponding mask, and the output is of the same resolution, *i.e.* no pre-processing to roll and pad or post-processing to upsample and blend is done.

We compare to the following related approaches:
- LaMa [43]: a ResNet-like inpainting network that relies on fast Fourier convolutions [5] to handle large masks;
- LGPN-Net [13]: estimates the room layout via Horizon-Net [42] and uses its edges to guide a GAN for inpainting;
- Stable Diffusion 2.0 inpainting [36].

Note that masks are dilated by 10-20 pixels for these methods, as they are not specifically trained to handle mask inaccuracies like ours, and respectively their results were worse without the dilation. No dilation is applied for our method.

Qualitative comparison We test on furnished panoramas from the Habitat dataset [35]. Figure 6 compares the results on a few examples. The ResNet- and GAN-based approaches, LaMa and LGPN-Net, tend to generate very blurry results. Moreover, they leave pieces of furniture objects un-inpainted. This is an even more prevalent effect with smaller or no mask dilation, as shown in the supplementary document. On the other hand, SD and our custom modification of it generate much crisper textures, even in very complex scenes. While plain SD inpainting hallucinates objects, like the tables in the first example, our method does not. In addition, SD leaves shadows behind as in the second example. The training of our method enables it to modify pixels outside of the inpainting mask, which is why it manages to remove the shadow and inpaint it with plausible floor texture.

Quantitative comparison To quantitatively compare results, we gather a selection of unfurnished panoramas from the Habitat dataset [35], insert synthetic furniture such as tables and ottomans on the floors, and measure the difference between the defurnished images produced by each method and the original unfurnished images. Note that the spaces produced this way have simpler furniture setups than in the
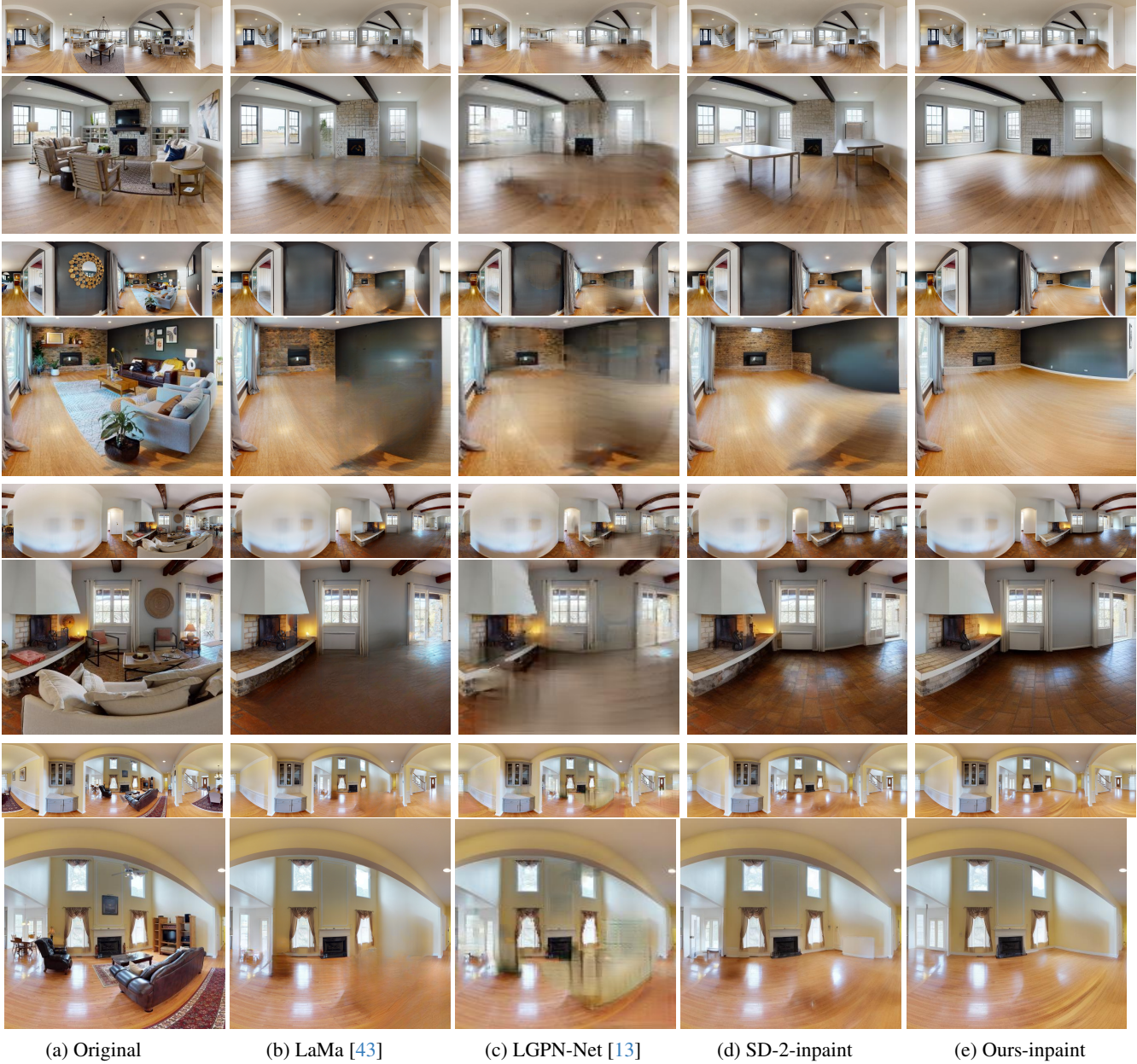
Figure 6. **Defurnishing comparison.** Examples are arranged in row pairs of full images and zoomed-in patches. Results from GAN-based approaches, LaMa and LGPN-Net, are blurry and contain remnants of furniture objects, even though inpainting masks had to be dilated for these methods to run optimally. SD generates crisper images, but may seek to explain shadows by hallucinating furniture. Our custom fine-tuning removes the hallucinations, resulting in coherent textures, even though it is the only one that does not require mask dilation.

natural images we used in the previous section. We evaluate absolute differences via PSNR, and perceptual differences via SSIM [48], LPIPS [55], and FovVideoVDP [31], a kind of just-objectionable-difference (JOD) measure for wide field-of-view images or video. Here, in addition to our inpainting network (Ours-inpaint), we evaluate our full pipeline (Ours-full), for a fairer comparison to a complete defurnishing method like LGPN-Net. The results are summarized in Table 1. As the synthetic furniture does not include rugs and leaves large areas from the original flooring visible (shown in supplementary document), LaMa and LGPN-Net manage to propagate these textures well into the inpainted regions, and thus achieve good PSNR and SSIM scores. However, the borders between inpainted and unpainted regions are blurry and disjoint, leaving these methods with low perceptual scores. On the other hand, our

(a) Original   (b) SD-2-inpaint, no mask dilation   (c) SD-2-inpaint, dilated mask   (d) Ours, no mask dilation

Figure 7. **LoRA fine-tuning.** Comparing off-the-shelf weights for Stable Diffusion v2 inpainting and our fine-tuned weights.

(a) Original; mask overlaid in blue   (b) Generated image   (c) Naïve replacement   (d) Our blending

Figure 8. **Blending.** Our blending technique combines the original and generated images according to the inpainting mask, shown in blue.

method and plain SD have somewhat lower-resolution textures, leading to worse PSNR. Conversely, they generate continuous textures with higher LPIPS and JOD. The post-processing component of our pipeline is developed exactly to remedy the low-resolution imagery—indeed, its better scores across all metrics demonstrate that we have accomplished this goal.

4.2. Ablation studies

Figure 7 shows an example inpainting result using off-the-shelf SD weights, with and without mask dilation. Note that without any mask dilation, the network hallucinates objects due to imperfect semantic segmentation and context from shadows. If the mask is generously dilated, the network does not have enough context to generate realistic inpaints for the kitchen island. Therefore, it is necessary to fine-tune SD on a custom dataset to generate inpaints of unfurnished spaces without requiring mask dilation.

Figure 8 shows how the image generated by our in-

painting network is combined with the original image according to the inpainting mask. The generated image is low-resolution and often contains undesirable artifacts (*e.g.* there are no trees through the windows of 8b), so we would prefer to use pixels of the original image when possible. However, naïvely replacing the generated image into the original image according to the inpainting mask (*i.e.* result = original·(1−mask)+generated·mask) creates visually jarring outlines due to physical shadows and differences in white-balance. While it is difficult to completely reconcile the differences between the detail frequencies of the original and generated images, our blending technique smoothly combines the two images.

Figure 9 demonstrates the effects of our rolling and padding pre-processing steps. As the goal of these steps is to ensure that inpaints are consistent across the "seam" of the panorama image, we show the crops at opposite edges of the panorama joined together, so the panorama seam lies in the center of the image. With no rolling, we obtain a ver-

(a) Original   (b) No rolling   (c) No padding   (d) With rolling and padding

Figure 9. **Rolling and Padding.** Inpainting results at the panorama seam with and without certain pre-processing steps.

(a) Original   (b) Inpainted

Figure 10. **Failure cases.** Structural changes or hallucinations may occur.

tical discontinuity where the seam lies. With no padding, we obtain low-quality inpaints due to the reduced context. While rolling and padding do not guarantee consistent inpainting around the entire panorama (*e.g.* this may fail when the mask spans the entire image), these two simple processing steps cover most cases.

4.3. Limitations and future work

While our method represents a step forward in defurnishing, it has limitations. Structural alterations and lingering hallucinations may occur, like the bench beneath the window in Figure 10. Our training methodology, which involves rendering synthetic furniture into unfurnished panoramas, introduces potential domain shift issues. This discrepancy between synthetic and real-world data may impact the quality of results. While we fixed the number of training prompts based on empirical evidence, the optimal prompting strategy remains an open question. Augmenting the fixed inference prompt "empty room" with additional contextual information could improve content generation accuracy.

Our reliance on the Stable Diffusion 2.0 inpainting architecture imposes constraints, necessitating modifications for higher resolution training and inference. The resolution of the output of Stable Diffusion model is too low - only 512 pixels in height, and the superresolution model sometimes made the output warped or unnatural. One area of improvement would be to choose an optimal superresolution model or to train one specifically on equirectangular images.

The applicability of our strategy to other generative models or imagery types besides indoor panoramas is out of the scope of this work and remains open for future research.

Finally, our method lacks consideration for multi-view consistency, which is important for applications like digital twins; each panorama is inpainted independently, overlooking valuable contextual information from other views. Future research should explore efficient and scalable approaches to incorporate geometric priors into the inpainting pipeline [54]. Being able to inpaint consistently across different views of the same underlying physical space may help with defurnishing other representations of digital twins, such as textured polygon meshes.

5. Conclusion

We have presented a novel approach to panorama defurnishing by utilizing domain-specific fine-tuning for Stable Diffusion inpainting. We have observed a notable reduction in undesirable hallucinations and improved the model's robustness to imperfect segmentation masks, making it much less likely to explain away shadows, light beams, reflections, and other similar effects. When compared to existing approaches, our method produces higher-quality results without the need for room layout estimation, as indicated in both qualitative and quantitative comparisons.

Acknowledgements

References

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B. Goldman. Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM SIGGRAPH 2009 papers*, 2009. 2

[2] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, page 417–424, USA, 2000. ACM Press/Addison-Wesley Publishing Co. 2

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 4

[4] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision Transformer Adapter for Dense Predictions. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 2, 3

[5] Lu Chi, Borui Jiang, and Yadong Mu. Fast Fourier Convolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 5

[6] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Object removal by exemplar-based inpainting. pages 721–728, 2003. 2

[7] Bruno Daniotti, Gabriele Masera, Cecilia Maria Bolognesi, Sonia Lupica Spagnolo, Alberto Pavan, Giuliana Iannaccone, Martina Signorini, Simone Ciuffreda, Claudio Mirarchi, Meherun Lucky, et al. The development of a bim-based interoperable toolkit for efficient renovation in buildings: From bim to digital twin. *Buildings*, 12(2):231, 2022. 1

[8] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A Universe of Annotated 3D Objects, 2022. 4

[9] Juan Manuel Davila Delgado and Lukumon Oyedele. Digital twins for the built environment: learning from conceptual and process models in manufacturing. *Advanced Engineering Informatics*, 49:101332, 2021. 1

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, pages 8780–8794. Curran Associates, Inc., 2021. 1, 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[12] Jonathan Eyre and Chris Freeman. Immersive applications of industrial digital twins. *The Industrial Track of EuroVR 2018*, pages 11–13, 2018. 1

[13] Chao-Chen Gao, Cheng-Hsiu Chen, Jheng-Wei Su, and Hung-Kuo Chu. Layout-guided Indoor Panorama Inpainting with Plane-aware Normalization. In *Asian Conference on Computer Vision (ACCV)*, 2022. 2, 5, 6

[14] V. Gkitsas, V. Sterzentsenko, N. Zioulis, G. Albanis, and D. Zarpalas. Panodr: Spherical panorama diminished reality for indoor scenes, 2021. 1, 2

[15] Vasileios Gkitsas, Nikolaos Zioulis, Vladimiros Sterzentsenko, Alexandros Doumanoglou, and Dimitrios Zarpalas. Towards full-to-empty room generation with structure-aware feature encoding and soft semantic region-adaptive normalization, 2021. 2

[16] Esa Halmetoja. The role of digital twins and their application for the built environment. *Industry 4.0 for the built environment: Methodologies, technologies and skills*, pages 415–442, 2022. 1

[17] James Hays and Alexei A. Efros. Scene completion using millions of photographs. *ACM Trans. Graph.*, 26(3):4–es, 2007. 2

[18] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. 3, 4

[19] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36:1–14, 2017. 2

[20] Isogawa, Mariko and Mikami, Dan and Iwai, Daisuke and Kimata, Hideaki and Sato, Kosuke. Mask Optimization for Image Inpainting. *IEEE Access*, 2018. 2

[21] Guanzhou Ji, Azadeh O Sawyer, and Srinivasa G Narasimhan. Virtual Home Staging: Inverse Rendering and Editing an Indoor Panorama under Natural Illumination. In *International Symposium on Visual Computing*, 2023. 2, 3

[22] Saleh Kalantari and Jun Rong Jeffrey Neo. Virtual environments for design research: Lessons learned from use of fully immersive virtual reality in interior design research. *Journal of Interior Design*, 45(3):27–42, 2020. 1

[23] 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. *arXiv:2304.02643*, 2023. 3

[24] Wolfgang Kuehn. Digital twins for decision making in complex production and logistic enterprises. *International Journal of Design & Nature and Ecodynamics*, 13(3):260–271, 2018. 1

[25] Prakhar Kulshreshtha, Nektarios Lianos, Brian Pugh, and Salma Jiddi. Layout Aware Inpainting for Automated Furniture Removal in Indoor Scenes. In *IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 2022. 2

[26] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. Planercnn: 3d plane detection and reconstruction from a single image, 2019. 2

[27] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions, 2018. 2

[28] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A Survey on Hallucination in Large Vision-Language Models, 2024. 1, 2, 3

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2015. 2

[30] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using Denoising Diffusion Probabilistic Models, 2022. 2

[31] Rafał K. Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. FovVideoVDP: a visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (SIGGRAPH)*, 40(4), 2021. 6

[32] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019. 2

[33] Stanley Osher, Martin Burger, Donald Goldfarb, Jinjun Xu, and Wotao Yin. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling & Simulation*, 4(2):460–489, 2005. 2

[34] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting, 2016. 2

[35] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 4, 5

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 2, 5

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 2

[38] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 2

[40] Muhammad Shahzad, Muhammad Tariq Shafiq, Dean Douglas, and Mohamad Kassem. Digital twins in built environments: an investigation of the characteristics, applications, and challenges. *Buildings*, 12(2):120, 2022. 1

[41] Yuhang Song, Chao Yang, Yeji Shen, Peng Wang, Qin Huang, and C. C. Jay Kuo. Spg-net: Segmentation prediction and guidance network for image inpainting, 2018. 2

[42] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. HorizonNet: Learning Room Layout With 1D Representation and Pano Stretch Data Augmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 5

[43] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust Large Mask Inpainting with Fourier Convolutions. *Winter Conference on Applications of Computer Vision (WACV)*, 2022. 2, 5, 6

[44] Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of Graphics Tools*, 9, 2004. 2

[45] S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models, 2024. 2

[46] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. In *International Conference on Computer Vision Workshops (ICCVW)*, 2021. 5

[47] YT Wang, C Liang, N Huai, J Chen, and CJ Zhang. A survey of personalized interior design. In *Computer Graphics Forum*. Wiley Online Library, 2023. 1

[48] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13 (4), 2004. 6

[49] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is Inevitable: An Innate Limitation of Large Language Models, 2024. 2, 3

[50] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention, 2018. 2

[51] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. Free-form image inpainting with gated convolution, 2019. 2

[52] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling, 2020. 2

[53] Edward Zhang, Ricardo Martin-Brualla, Janne Kontkanen, and Brian Curless. No Shadow Left Behind: Removing Objects and their Shadows using Approximate Lighting and Geometry. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[54] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2, 8

[55] 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. 6

[56] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. PanoContext: A whole-room 3D context model for panoramic scene understanding. In *European Conference on Computer Vision (ECCV) oral presentation*, 2014. 2

[57] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network, 2017. 2

[58] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene Parsing through ADE20K Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3

[59] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2