A. Appendix/Supplemental material

A.1. Intermediate Results from MVMD Blocks

To illustrate the functionality of each network component, we provide visualizations of attention heat maps and intermediate results. These visualizations demonstrate how the network processes input images, extracts information using attention mechanisms, and refines the mirror mask for accurate detection.

As discussed in Section 4, our MVMD network takes three images from a single scene as inputs. Fig. 1 shows an example of an indoor bathroom scene with a large mirror and complex textures, which challenges mirror detection by making it difficult to distinguish the reflection area from the surrounding wall.

The Inter-Views Block processes image pairs $[I_1, I_2]$ and $[I_1, I_3]$ through cross-attention and self-attention mechanisms. Fig. 2 visualizes the attention heat maps, revealing that cross-attention focuses on the left side of the mirror where reflection differences between $[I_1, I_2]$ and $[I_1, I_3]$ are evident. The differences in their cross-attention heatmaps indicate that each pair captures distinct mirror information due to viewpoint shifts and mirror reflections. Thus, using two pairs ensures that sufficient information is captured through these shifts. As discussed in Section 5.4, this shows the Inter-Views Block's ability to capture object shifts inside the mirror due to viewpoint changes. Fig. 3 depicts the output of the Inter-Views Block after channel attention. It shows that unimportant feature channels are filtered out (visible in the upper-left corner), while significant channels are retained, as demonstrated in the remaining figures.

As discussed in Section 5.4, the Intra-view Block captures the correspondence between objects inside and outside the mirror. Fig. 4 illustrates the output feature channels, highlighting how reflected objects inside the mirror match real objects outside. The same channel attention mechanism is applied to filter out less important channels while preserving key ones, ensuring precise mirror location information.

Finally, the Refinement Block enhances the edge definition of the initial mask generated by the Inter-View and Intra-view Blocks. Fig. 5 shows how it significantly improves mirror edge clarity and corrects misidentified mirror areas, particularly around objects in front of the mirror, such as the light bulbs at the top left and right.

A.2. Additional MVMD Results

To demonstrate the accuracy and robustness of the proposed MVMD module across a variety of scenes and its stability in detecting mirrors from multiple viewpoints within the same scene, we present additional inference results in Fig. 6. For each scene, three images were selected from different angles of real-world environments. These results demonstrate the module's high accuracy in detecting mirrors, even under challenging conditions such as indistinct edges and complex environments. Furthermore, they illustrate the module's consistent performance across different perspectives, showcasing its ability to reliably identify and delineate mirrors regardless of viewpoint changes. This comprehensive evaluation underscores the effectiveness of the MVMD module in handling diverse and complex mirror detection scenarios.

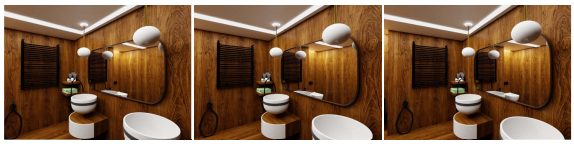

Figure 1. Example scene with three input images: I_1, I_2, and I_3 in an indoor bathroom.

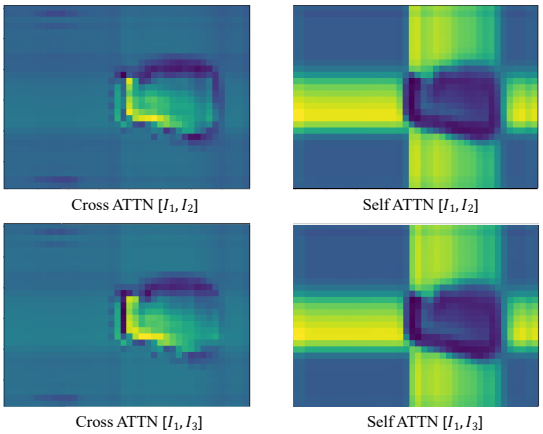

Figure 2. Visualization of the cross-attention and self-attention heat maps on $[I_1, I_2]$ and $[I_1, I_3]$ in the Inter-views Block.

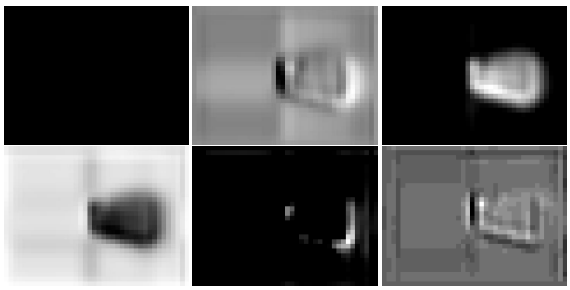

Figure 3. Visualization of selected channels from the output of the Inter-views Block.

A.3. Revisit MVMD Dataset

To highlight the diversity of our MVMD dataset, particularly in terms of mirror shapes, sizes, locations, and

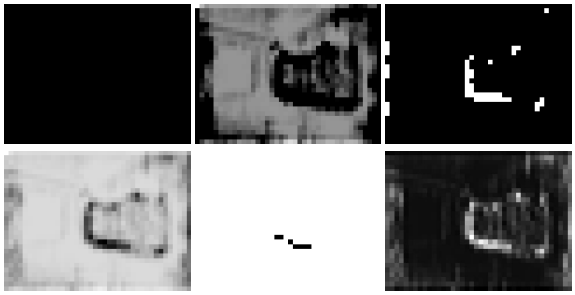

Figure 4. Visualization of selected channels from the output of the Intra-view Block.

Figure 5. Visualization of the differences between the initial mask and final mask.

quantities, we provide additional examples in Fig. 7. These examples showcase a wide range of mirror configurations within various scenes, illustrating the dataset's capacity to cover different mirror types and placements. The variety in mirror attributes, such as large versus small mirrors, different shapes, and their diverse positions within the scene, underscores the comprehensive nature of the dataset. By including these varied examples, we demonstrate how the dataset captures the real-world complexity of mirror detection, which is crucial for training and evaluating the MVMD network under diverse conditions.

A.4. From Mirror Mask to 3D Labels

Various methods in 3D reconstruction use different types of image masks to modify scenes or enhance reconstruction quality based on specific tasks [8, 9, 11]. Image masks are particularly crucial in challenging areas such as mirrors, where they significantly improve accuracy [5, 6, 12]. However, these masks are often created manually, a process that is labor-intensive and prone to errors, especially around complex edges. Such inaccuracies can result in sub-optimal reconstructions and artifacts in specific viewpoints.

To address these issues, we propose the MVMD network for automatically generating mirror masks, tailored to match the input requirements of 3D reconstruction tasks that use multi-view inputs. Our approach begins by generating a binary mask based on pixel-wise confidence scores from 2D images. This data-driven technique enhances the precision of mirror area identification, reducing manual errors and inconsistencies, and thus improving both reconstruction accuracy and efficiency.

Furthermore, to effectively utilize the automatically generated mirror mask, it can be mapped into 3D space using established labeling and voting algorithms, commonly employed in traditional 3D point cloud reconstructions. Labeling algorithms assign semantic categories to each point in the point cloud, aiding in scene classification and segmentation. Voting algorithms, such as Hough Voting [10], RANSAC [2], and kNN voting [3], leverage global information from local features to enhance point cloud denoising, plane fitting, and object detection.

For NeRF (Neural Radiance Fields) [7] and 3D Gaussian Splatting (3D GS) [4], labeling and voting techniques can still be adapted, despite their different representations of 3D data compared to traditional point clouds. NeRF generates a continuous field using neural networks rather than explicit point clouds. Adaptations such as Semantic NeRF [1] extend the model to output semantic information for scene annotation. Similarly, the geometric structures produced by 3D GS can be converted into discrete point clouds [4], facilitating the use of traditional labeling techniques for classification and segmentation. Correspondingly, voting techniques can be applied indirectly to NeRF by first extracting point clouds from the NeRF output and then using these techniques to enhance segmentation or denoising. In the case of 3D GS, voting can be applied directly to point clouds or meshes obtained after the Structure from Motion (SfM) process, improving reconstruction accuracy and maintaining consistent object segmentation across multiple views.

By integrating these labeling and voting techniques, the proposed MVMD network effectively enhances the reconstruction accuracy of mirror regions in various 3D reconstruction tasks. It automatically generates mirror masks and maps them into 3D space, combining labeling with classification to precisely capture mirror regions and manage reflection effects. This approach reduces geometric errors and confusion caused by reflections, significantly improving the overall fidelity and reliability of 3D scene reconstruction. To further validate the effectiveness of the MVMD network, we plan to conduct a system-level integration and perform extensive experiments in future work to comprehensively assess its performance and potential applications in various complex scenarios. This will help to further optimize the model and ensure its reliability and effectiveness in practical applications.

A.5. Extending to High-Reflectivity Regions

In the real world, objects with high-reflective surfaces often exhibit characteristics similar to mirrors, which poses significant challenges for 3D reconstruction. High-reflectivity textures can mislead algorithms into interpreting reflections as part of the texture or as actual objects, lead-

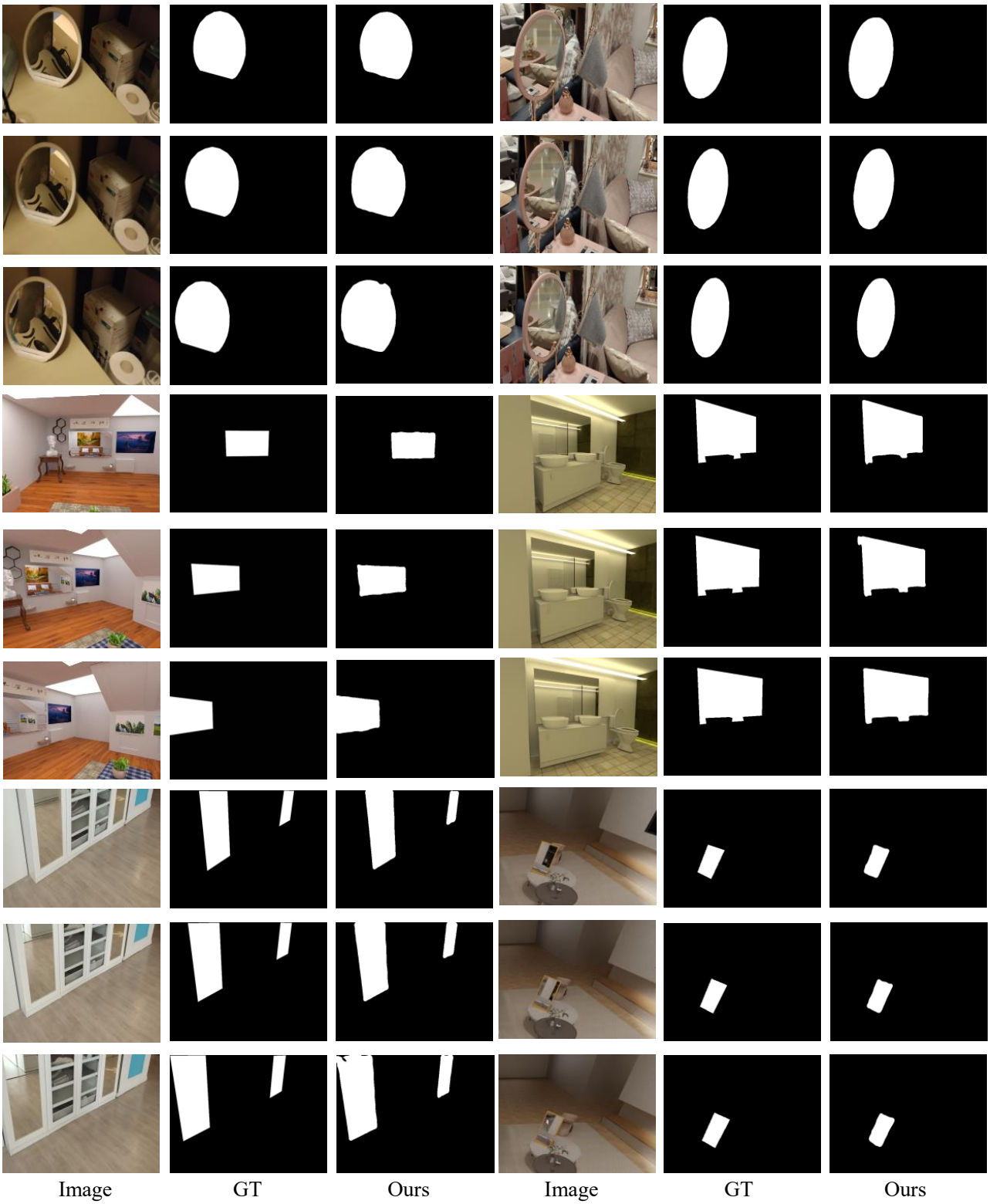

Figure 6. Additional MVMD network results on different scenes, featuring three viewpoints per example.

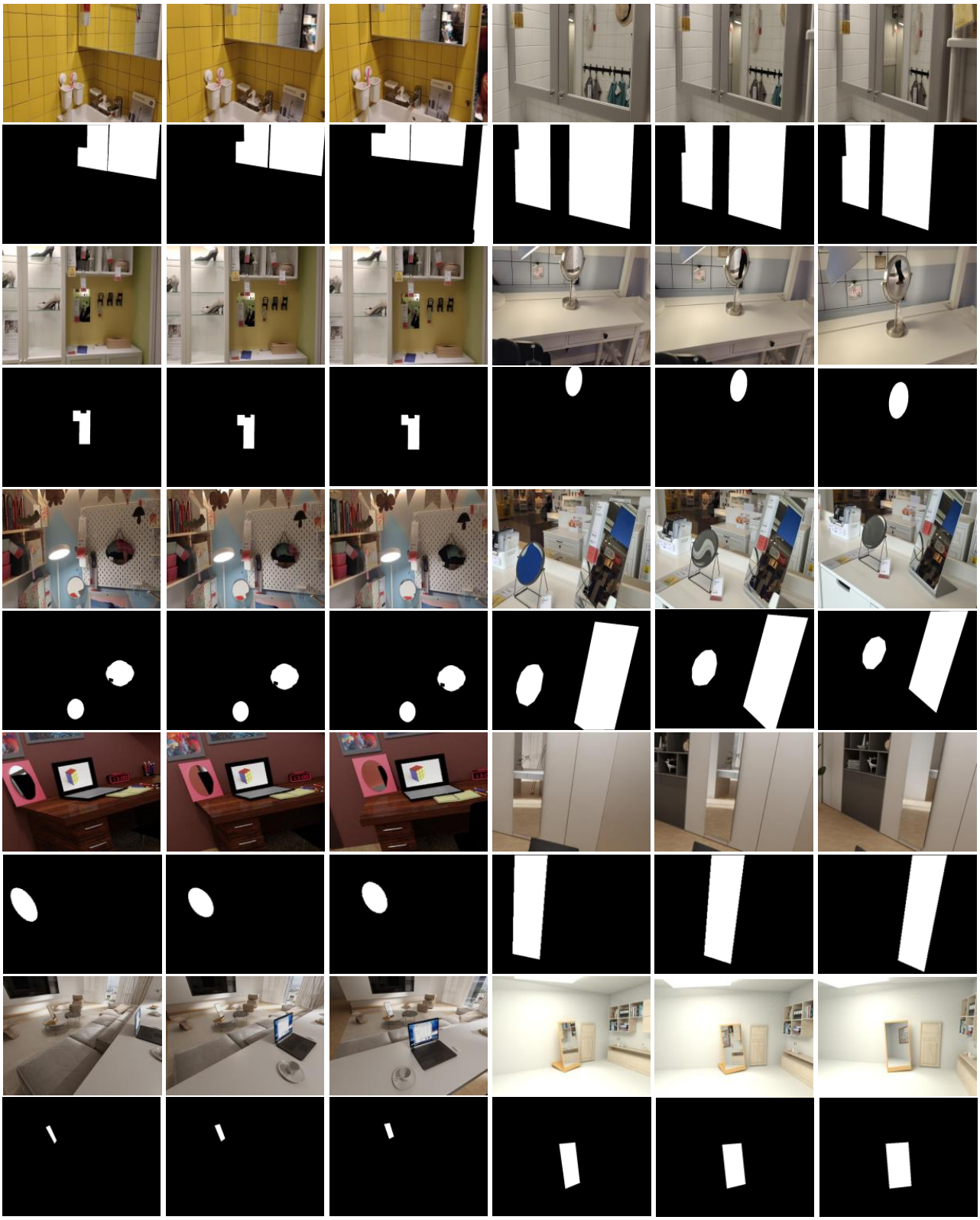

Figure 7. Additional examples from the proposed MVMD dataset.

ing to artifacts such as ghosting or errors depending on the viewing angle.

Directly applying mirror detection networks to identify all high-reflectivity areas remains challenging. Firstly, high-reflectivity surfaces exhibit greater variability compared to mirrors, as these surfaces are often uneven, leading to distortion or blurring in reflected content. Secondly, unlike mirrors with near-perfect reflections, high-reflectivity areas often display content influenced by the original texture and reflection coefficient, which can result in reflected external objects appearing as part of the surface. Finally, high-reflectivity areas may lack distinct edges, complicating the task of delineating actual object boundaries.

However, by leveraging the cross-attention and self-attention mechanisms of our MVMD network, there is potential to extend its capabilities to detect high-reflectivity regions. Specifically, the basic structure of our Inter-views Block can be retained to differentiate high-reflectivity areas from the background, as it is adept at identifying varying levels of reflection. For the Intra-view Block, we can adapt its design to handle the effects of original texture and reflection coefficients, enabling it to generate an initial mask for high-reflectivity regions. Finally, by enhancing the edge detection capabilities of the Refinement Block, we can improve the network's ability to distinguish texture-based edges within these high-reflectivity areas, refining the mask to better represent the actual boundaries of objects.

In future work, we plan to systematically evaluate and refine these extensions through extensive experimentation with various high-reflectivity scenarios. Our goal is to enhance the MVMD network's robustness and accuracy in detecting and delineating high-reflectivity regions. By doing so, we aim to improve its applicability across a broader range of real-world environments, supporting next-generation 3D reconstruction applications.

References

[1] Yuedong Chen, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Sem2nerf: Converting single-view semantic masks to neural radiance fields. *arXiv preprint arXiv:2203.10821*, 2022. 2

[2] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[3] Jianping Gou, Taisong Xiong, Yin Kuang, et al. A novel weighted voting for k-nearest neighbor rule. *J. Comput.*, 6(5):833–840, 2011. 2

[4] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. 2

[5] Jiayue Liu, Xiao Tang, Freeman Cheng, Roy Yang, Zhihao Li, Jianzhuang Liu, Yi Huang, Jiaqi Lin, Shiyong Liu, Xiaofei Wu, et al. Mirrorgaussian: Reflecting 3d gaussians for reconstructing mirror reflections. *arXiv preprint arXiv:2405.11921*, 2024. 2

[6] Jiarui Meng, Haijie Li, Yanmin Wu, Qiankun Gao, Shuzhou Yang, Jian Zhang, and Siwei Ma. Mirror-3dgs: Incorporating mirror reflections into 3d gaussian splatting. *arXiv preprint arXiv:2404.01168*, 2024. 2

[7] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2

[8] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Marcus A. Brubaker, Jonathan Kelly, Alex Levinshtein, Konstantinos G. Derpanis, and Igor Gilitschenski. Reference-guided controllable inpainting of neural radiance fields. In *ICCV*, 2023. 2

[9] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G. Derpanis, Jonathan Kelly, Marcus A. Brubaker, Igor Gilitschenski, and Alex Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20669–20679, 2022. 2

[10] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 2

[11] Silvan Weder, Guillermo Garcia-Hernando, Áron Monszpart, Marc Pollefeys, Gabriel Brostow, Michael Firman, and Sara Vicente. Removing objects from neural radiance fields. In *CVPR*, 2023. 2

[12] Junyi Zeng, Chong Bao, Rui Chen, Zilong Dong, Guofeng Zhang, Hujun Bao, and Zhaopeng Cui. Mirror-nerf: Learning neural radiance fields for mirrors with whitted-style ray tracing. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4606–4615, 2023. 2