Digital Environment Description and Reconstruction using Panoptic Segmentation

João Francisco de Souza Santos Lemos[1[0009-0005-1244-7656]], Gabriel Amaral Dorneles[1[0009-0001-6554-5832]], Igor Pardo Maurell[1[0000-0002-4376-9544]], Stephanie Loi Brião[1[0000-0001-9345-2038]], Rodrigo da Silva Guerra[1[0000-0003-4011-0901]], and Paulo Lilles Jorge Drews Junior[1[0000-0002-7519-0502]]

Center for Computational Science - C3,
Universidade Federal do Rio Grande - FURG, Brazil
jfsslemos@gmail.com, gdorneles@furg.br
http://www.c3.furg.br

Abstract. Simulation in the field of robotics is a less costly, safer, more scalable, and more versatile alternative to real-world scenarios. In this work, we present a pipeline for the digital reconstruction of environments using the panoptic segmentation technique. The overall aim is to automate the transcription of real geometries into the digital environment, creating a virtual scene where objects are meticulously segmented, making it possible to manipulate, replicate, remove, and practically move these objects. The methodology is based on Neural Radiance Field for segmentation and reconstruction of scenes, classical methods for manipulating structures, SDFormat as a scene description file, and Gazebo as a 3D robotic simulation platform. The results show that the method is capable of transcribing the scene into a virtual environment and using it as a space for testing, validating, and training robots.

Keywords: panoptic segmentation · simulated environment · surface reconstruction.

1 Introduction

Simulation in robotics accelerates the design process, enables detailed testing, and is more cost-effective than real-world experimentation [2]. It provides exhaustive repetition of scenarios, exploring different conditions, and precisely adjusting parameters, something often unfeasible or costly in the physical world. It also facilitates controlled testing without posing risks to the physical safety of agents or causing damage to real-world environments. However, one of its main challenges is the difficulty of constructing scenes that accurately mimic the real world, especially considering time-critical environments such as those presented in the Robocup@Home competition.

The method described in this paper allows for the acquisition of a digital representation of the Robocup competition arena, allowing teams to simulate different tasks without manually creating the digital environment, a process that can be very challenging and time consuming [14].

This paper presents a novel digitalization pipeline using panoptic segmentation to achieve faithful digital representations of real-world environments, with posed images as input. Panoptic segmentation has been increasingly utilized in the literature to generate a meaningful representations environments [5, 18]. Building upon these results, this research aims to create a tool that can reconstruct a real environment into a virtual scene. In this virtual setting, objects are meticulously segmented, enabling straightforward manipulation, replication, removal, and relocation of these objects.

Our pipeline employs posed images from a given scene and a method based on DM-NeRF [15] for segmentation and reconstruction of the environment with no prior object detection or segmentation. The resulting three-dimensional meshes are then converted into XML files, categorized according to their panoptic classification, to produce an SDF file compatible with simulation software. A depiction of the proposed method can be seen in Fig. 1. The method is evaluated on the DM-SR [15] dataset and we integrate the obtained results into the Gazebo simulation environment through ROS.

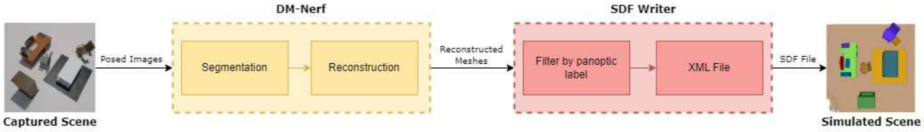


Fig. 1: The proposed pipeline: All individual objects are segmented and reconstructed with the appropriate labels, then transformed into the required format for simulation.

2 Related Work

This section provides an overview of existing research and developments relevant to this work, to contextualize the diversity of methodologies and the progression of ideas in the subject area. The topics explored include panoptic segmentation and reconstruction techniques and 3D scene reconstruction methods.

Image Segmentation: Image segmentation is a fundamental task in computer vision, aiming to partition an image into multiple segments or regions of interest, allowing various downstream tasks such as object detection, recognition, and scene understanding.

Semantic segmentation [3] is a classification where each pixel in an image is assigned a class label representing the category it belongs. Semantic segmentation provides a granular understanding of the image by labeling each pixel.

However, instance segmentation [9] takes semantic segmentation a step further by not only assigning class labels to each pixel but also distinguishing between individual object instances. This result is achieved by only categorizing the pixels that are given as relevant in the image, not guaranteeing that all pixels will be classified.

Panoptic Image Segmentation: Panoptic segmentation [6] goal is to unify semantic and instance segmentation into a single task that can segment all pixels in an image, regardless of whether they belong to objects or stuff (e.g., sky, grass, road). In panoptic segmentation, each pixel is labeled with a class label if it belongs to a thing (object) or stuff (background). This holistic approach is crucial for tasks requiring a complete understanding of the visual environment, such as scene understanding and content-aware robotic tasks.

3D Scene Reconstruction: Several methodologies for 3D scene reconstruction from RGB images have been explored in prior research. Dahnert et. al [5] produces results from a single view, so it encounters limitations in reproducing a full scene without employing a mesh registration method. Uni-3d [18] utilizes a multi-view approach but, similarly to Dahnert et. al, it relies on traditional reconstruction techniques, which leads to a compromise in the fidelity of the real-world geometry represented. Atlas [12] attempts to address these issues; however, its efficiency is tied to a segmentation method that is purely semantic, resulting in objects lacking specific instance-based information. Contrasting these approaches, this paper employs the DM-NeRF [15] method, which is among the first to include implicit representations of all 3D objects in complex scenes, relying solely on color images and 2D object labels for supervision.

3 Method

In this section, we detail the development of the pipeline designed to digitally recreate a real scene. Fig. 1 illustrates the system's architecture, depicting a sequence of four main components along with their respective internal elements.

3.1 Panoptic Segmentation and Reconstruction

Panoptic segmentation [10] unifies semantic and instance segmentation into a single result, assigning semantic labels and indices to all pixels in an image. In contrast to traditional instance segmentation, which employs the concept of relevant classes to decide if a pixel should be classified, panoptic segmentation assigns semantic labels to all pixels, but does not provide specific indices for those that belong to irrelevant classes.

As discussed in Sec. 2, segmentation and reconstruction are well-researched fields with a variety of developed methods, but they are not easily integrated. To connect the two, the segmentation results must be handled by transcribing the data format, filtering the segmentation masks, supplementing the data, and taking other steps to ensure reconstruction can occur.

To address this issue, the DM-NeRF method constructs a triangular mesh by retrieving structures from continuous 3D scenes, segments all objects individually in three-dimensional space, and allows for flexible manipulation of objects. Consequently, the segmentation and reconstruction processes occur simultaneously, since the vector predicted for each of the pixels by segmentation also corresponds to a specific color that is assigned to the vertices in the scene reconstruction.

3.2 Reconstruction Post-processing

Once the scene has been reconstructed, it is important to isolate the objects so that the vertices of each item are separated. This segregation is achieved through a filtering process that utilizes the panoptic label, divided into semantic type labels and instance index labels. These labels enable the distinction between objects of the same category (such as different chairs or tables in the scene, in a domestic environment), and through the semantic tag it is possible to write each object's type within the description file.

Due to the inherent inaccuracies in the reconstruction process, it fails to ensure uniformly smooth surfaces, resulting in uneven terrain. This impairs the navigation and stability of a robotic base. To circumvent this issue, a plane is estimated through primitive segmentation using RANSAC fitting [8]. To find the plane that represents the ground, the sub-mesh of vertices classified with the corresponding label is selected and used as input for the segment plane method of the consolidated Open3D library [13].

The method accepts three arguments: distance threshold θ_t defines the maximum distance a point can have to an estimated plane to be considered a point belonging to it; ransac n defines the number of points that are randomly sampled at each iteration to estimate a plane; and number of iterations λ defines how many times a random plane is sampled and checked. Subsequently, the function outputs a list of indices for points deemed to belong to the plane, alongside a description of the plane through its general equation.

Each vertex of the reconstructed mesh must be interpreted as a point, enabling the plane segmentation method to fine-tune parameters that most accurately describe the ground plane. As a result, each of the vertices has its three-dimensional coordinates projected onto the plane, maintaining the details of the mesh, but with a perfectly smooth floor, a process that is illustrated in Fig. 2.

Once the reconstruction is completed, the representation is in a different coordinate system than the simulator, which leads to mismatches in spatial orientation and positioning when integrating the reconstructed scene into the simulated environment. To address this issue, a coordinate system transformation process is necessary.

Fortunately, estimating a plane provides the information that makes up the normal vector, which is the exact information needed to adjust the orientation. Nonetheless, it is necessary to find the rotation matrix between the scene and

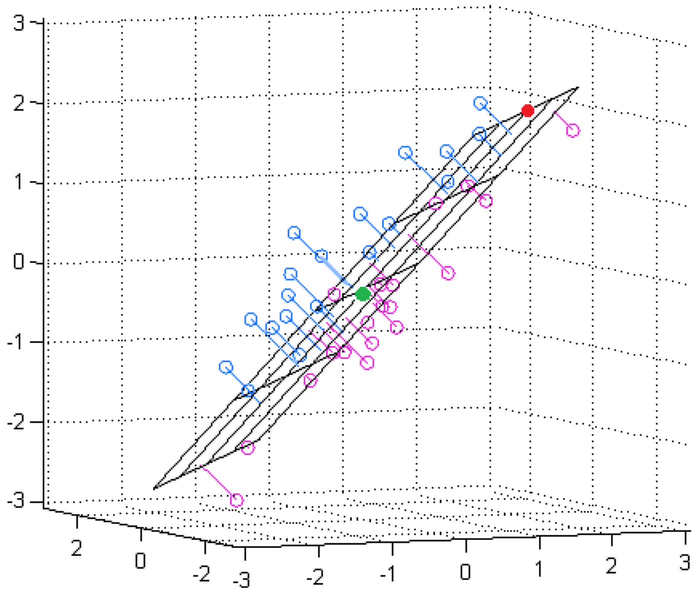


Fig. 2: Points being projected onto a predicted plane [1].

a line parallel to the z-axis. To find the rotation matrix θ between the normal vector $\mathbf{n} = [n_x, n_y, n_z]$ and the positive axis z^+, the scalar product is used:

$$\cos(\theta) = \hat{\mathbf{n}} \cdot [0, 0, 1] = \frac{n_z}{|\mathbf{n}|} \tag{1}$$

In which $|\mathbf{n}|$ is the magnitude of the normal vector and $\hat{\mathbf{n}}$ is the normalized vector. The angle can be calculated using the inverse cosine function.

Applying this angle to the orientation of the overall scene guarantees objects parallel to the z-axis, despite the normal vector not specifying whether the mesh is oriented upwards or downwards. To solve this problem, we determine the height of the mesh on both sides of the floor and orient the side with greater height to face upwards.

3.3 Setting up the Gazebo Scene

To rebuild a scene for robotics simulation, it is important to use a simulator that allows for robot control, sensor integration and provides a user-friendly interface. To ensure this operates smoothly, effective communication among the various components of the robotic system is crucial. The Gazebo simulator was chosen for its integration with ROS, which offers specific interfaces, known as plugins, that enable communication between Gazebo and robotic systems. These plugins, shown in Fig. 3, facilitate the exchange of information between the simulated environment and the various software components operating within the robot.

The Gazebo simulator receives an SDF (Simulation Description Format) file as an input. This file contains the complete description of the simulated environment. To write the file, the ElementTree python library [1] is used, which implements a simple and efficient interface for creating and editing XML data. The real world can be abstractly defined as a collection of objects and backgrounds that relate hierarchically in terms of belonging. This approach makes it possible to use SDF files to describe the scene, since it is essentially composed of labels and attributes in a tree-like structure.

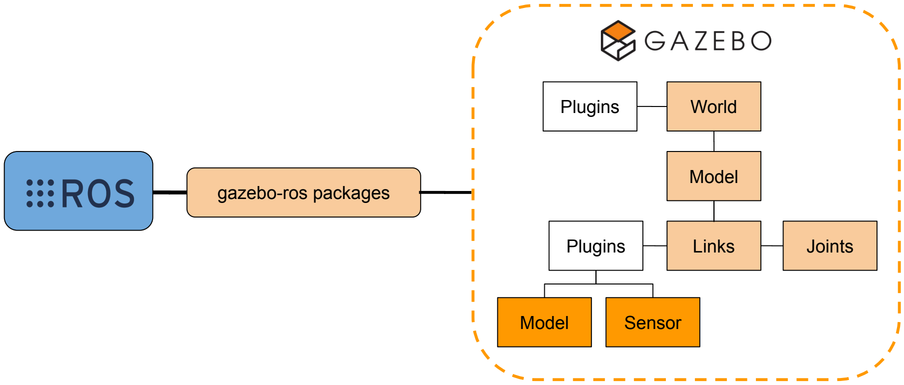


Fig. 3: Communication flow between ROS and Gazebo.

4 Evaluation

This chapter breaks down the results into separate sections, each focusing on the different analyses—quantitative, empirical, or patterns observed. We explain the impact of each process in detail to clarify how they contribute to the overall results. The experiments were designed to assess the effectiveness of the chosen methods and to better understand their performance in a simulated environment.

4.1 Validation conditions

As validation data for the pipeline, the DM-SR [15] was chosen, a *dataset* made up of residential environments designed precisely for quantitative evaluation of scene reconstruction and panoptic segmentation tasks. It already provides a multi-view capture of the scene with 95 “posed images”. Additionally, it includes

[1] ElementTree documentation: https://docs.python.org/3/library/xml.etree.elementtree.html.

semantic and instance annotations for all objects in each image, simplifying the quantification of training and testing metrics.

The dataset is made up of complex domestic scenes with dimensions close to 24 m^2 in area and 3 m in height, containing around 8 objects. We focused our study on the study room model, shown in Fig. 4a, selected due to its large number of elements, a variation of sizes and shapes, sparse positioning, and overlapping objects. With these characteristics, the model effectively addresses the main potential challenges presented by an indoor setting in domestic environments, though further investigation is required to address more specific conditions.

4.2 Panoptic Reconstruction

The overall quality of the panoptic reconstruction varies based on the results of two key internal processes of the method, view synthesis and decomposition. In computer graphics, view synthesis involves generating novel images of a specific scene when the only available information is photos taken from different viewpoints. Decomposition is the separation of the elements that make up a scene.

Using the model in the DM-NeRF network and evaluating it with the metrics proposed by Mildenhall et al. [11], the method described in this paper is capable of achieving significant results shown in Table 1.

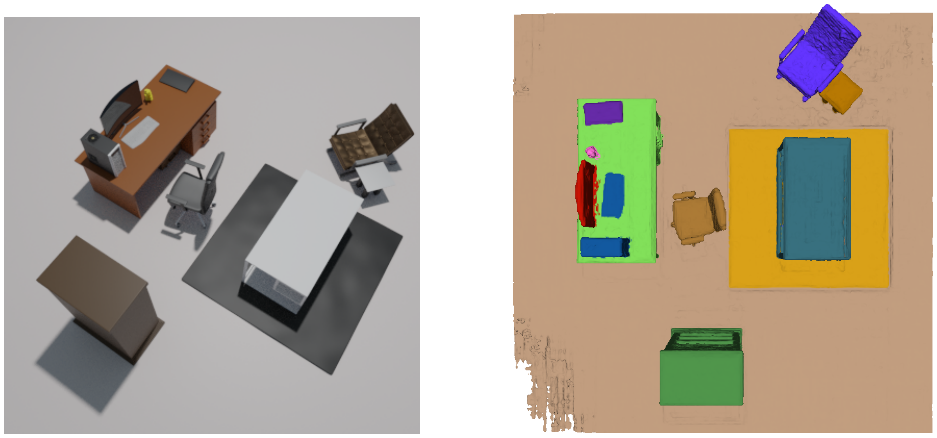
Scene	View synthesis			Decomposition (%)
	PSNR ↑	SSIM ↑	LPIPS ↓	
Bathroom	45.22	0.991	0.022	99.20
Bedroom	47.59	0.994	0.018	100.00
Dining Room	38.76	0.987	0.015	99.49
Kitchen	48.90	0.989	0.028	100.00
Reception	41.92	0.996	0.011	99.60
Living Room	43.74	0.981	0.024	100.00
Study Room	39.67	0.995	0.008	100.00
Office	47.01	0.992	0.012	99.12
Mean	44.66	0.991	0.017	**99.58**

Table 1: Results obtained from the respective metrics.

PSNR (Peak Signal-to-Noise Ratio) [7] is a metric that, in the case of RGB images, ranges from the worst case at 0 to the best possible result at 58.06. The average PSNR of 44.66 indicates a small difference or error between the original image and its reconstructed version, as PSNR quantifies the ratio between the maximum possible signal energy (original image) and the noise energy (error) affecting its fidelity.

The SSIM (Structural Similarity Index) [16], a metric that ranges from -1 to 1, shows an average SSIM of 0.991. This indicates that the final and reconstructed images are very close to being identical in terms of structure, luminance, and contrast.

The LPIPS (Learned Perceptual Image Patch Similarity) [17] coefficient is inversely proportional to similarity, so the value near 0.015 endorses the significant outcome of the previous metrics in terms of vision synthesis.

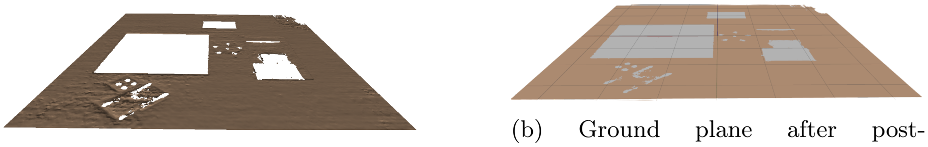


(a) “Study Room” environment from the DM-SR dataset [15].

(b) Results from the panoptic reconstruction of “Study Room”.

Fig. 4: Comparison between the original and reconstructed scene.

Finally, the average decomposition accuracy percentage being so close to 100 shows that the object vertices are correctly segmented with only minimal noise present. The high quality of the results is apparent when comparing the original scene, shown in Fig. 4a, with the reconstructed scene in Fig. 4b.



(a) Uneven ground plane.

(b) Ground plane after post-processing.

Fig. 5: Comparison between the ground plane before and after post-processing.

The segment plane method described in section 3.2 successfully transformed an uneven ground plane into a continuous plane that retained the mesh details, making it ideal for robot traversal.

In the reconstructed scene, the decomposition of the mesh allows the scene to be mutable, since the objects are not attached. The quality of the separation directly reflects the quality of the segmentation, with colors assigned according to the panoptic class attributed to each vertex, as depicted in Fig. 6.

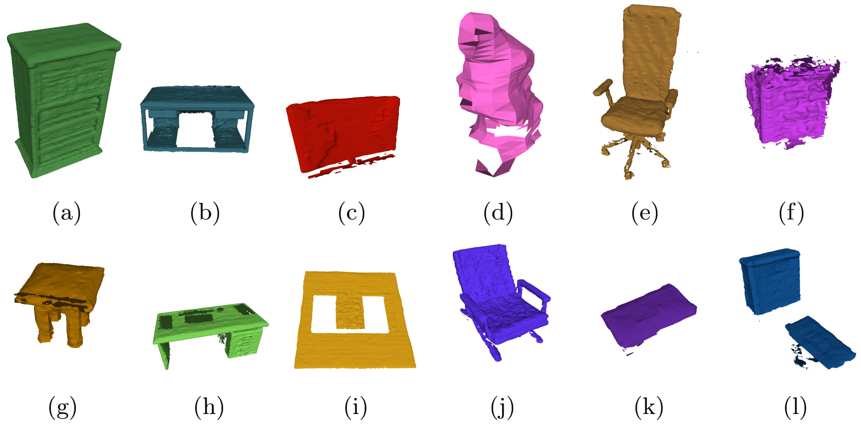


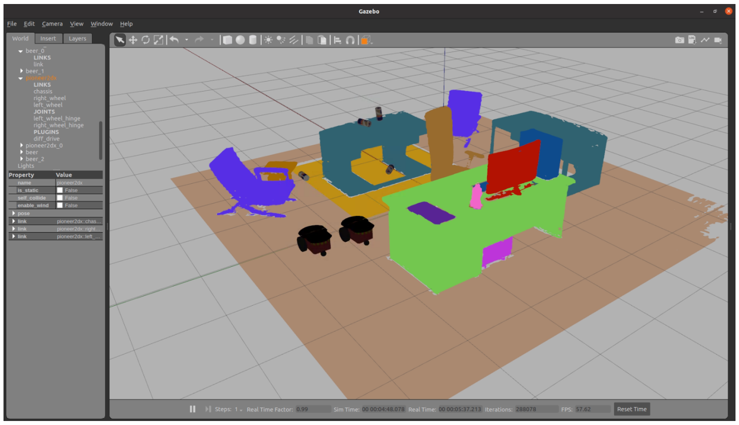
Fig. 6: Objects of the DM-SR dataset segmented and separated by their classes and instances.

4.3 Simulated Environment

As described in section 3.3, the algorithm developed to write the SDF file is simple. It creates a tree structure for the simulated world, assigning the visual characteristics and collision properties to each of the objects. If necessary, the scenario is rotated to align with the orientation of the simulator.

The world allows objects to be replicated, moved, rotated, or removed. The scenario is capable of incorporating other external elements that were not initially present and, as the main feature, serves as a platform that interacts with robots. All these features are displayed in Fig. 7, where new objects, such as cans, are added on top and beside the blue table. In this example, the cabinet is removed, the blue table is replicated, and two robotic bases are positioned in the scene, where they can navigate freely.

One of the applications for these features is the aforementioned Robocup@Home competition, where the limited time allotted for each team in the arena makes creating a simulated environment both challenging and valuable, while the controlled nature of the environment serves as an ideal setting. Using the described method, posed images of the arena can be captured to quickly generate a simulated version of all objects in the scene, allowing different tasks to be tested under conditions similar to the real-world environment, while easily modifying the layout for any specific requirements from different tests.



Property	Value
name	pioneer2dx
is_static	□ False
self_collide	□ False
enable_wind	□ False
pose	
link	pioneer2dx::chas...
link	pioneer2dx::right...
link	pioneer2dx::left_...

Steps: 1 Real Time Factor: 0.99 Sim Time: 00 00:04:48.078 Real Time: 00 00:05:37.213 Iterations: 288078 FPS: 57.62

Fig. 7: Simulated environment with newly added, removed, and replicated objects, alongside two robots positioned in the scene.

5 Conclusions

In this work, we presented a pipeline for the digital reconstruction of environments based on NERFs and using panoptic segmentation. The proposed method aims to create a virtual environment by tracing real-world geometries, allowing manipulation, replication, and removal of objects within the scene. Using Gazebo as a simulation environment demonstrates that the method can generate virtual robotic testing and training environments.

Quantitative evaluations of the scene reconstruction and segmentation processes yielded results with high fidelity and minimal noise. The pipeline was validated on the DM-SR dataset, which features eight typical household rooms. Our method produces accurate digital renditions of real-world environments and integrates them into the Gazebo simulation platform. This fact enables robots to navigate and interact within the virtual scene, allowing system validation without the risks and costs associated with real-world experiments.

Future work includes assessing the method across various datasets, including custom datasets based on real-world environments. This will help us investigate the method's efficiency in more specific scenarios, particularly those involving cluttered environments. Furthermore, state-of-the-art techniques based on retrieve and deform templates, such as those proposed by Dahnert et al. [4], should be explored to ensure more realistic structures despite incomplete scanning.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Azevedo, R.: Fitting a plane to points using svd. Mathematics Stack Exchange, https://math.stackexchange.com/q/3501135

2. Choi, H., Crump, C., Duriez, C., Elmquist, A., Hager, G.D., Han, D., Hearl, F., Hodgins, J.K., Jain, A., Leve, F.A., Li, C., Meier, F., Negrut, D., Righetti, L., Rodriguez, A., Tan, J., Trinkle, J.C.: On the use of simulation in robotics: Opportunities, challenges, and suggestions for moving forward. Proceedings of the National Academy of Sciences of the United States of America **118** (2020), https://api.semanticscholar.org/CorpusID:229281163
3. Csurka, G., Volpi, R., Chidlovskii, B.: Semantic image segmentation: Two decades of research (2023)
4. Dahnert, M., Dai, A., Guibas, L.J., Nießner, M.: Joint embedding of 3d scan and CAD objects. CoRR **abs/1908.06989** (2019), http://arxiv.org/abs/1908.06989
5. Dahnert, M., Hou, J., Nießner, M., Dai, A.: Panoptic 3d scene reconstruction from a single rgb image. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021)
6. Elharrouss, O., Al-Maadeed, S., Subramanian, N., Ottakath, N., Almaadeed, N., Himeur, Y.: Panoptic segmentation: A review (2021)
7. Fardo, F.A., Conforto, V.H., de Oliveira, F.C., Rodrigues, P.S.S.: A formal evaluation of PSNR as quality measurement parameter for image segmentation algorithms. CoRR **abs/1605.07116** (2016), http://arxiv.org/abs/1605.07116
8. Fischler, M.A., Bolles, R.C.: 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)
9. Hafiz, A.M., Bhat, G.M.: A survey on instance segmentation: state of the art. International Journal of Multimedia Information Retrieval **9**(3), 171–189 (Jul 2020). https://doi.org/10.1007/s13735-020-00195-x, http://dx.doi.org/10.1007/s13735-020-00195-x
10. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9404–9413 (2019)
11. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. CoRR **abs/2003.08934** (2020), https://arxiv.org/abs/2003.08934
12. Murez, Z., van As, T., Bartolozzi, J., Sinha, A., Badrinarayanan, V., Rabinovich, A.: Atlas: End-to-end 3d scene reconstruction from posed images (2020)
13. Open3D Contributors: Open3D: A modern library for 3d data processing. http://www.open3d.org (2024)
14. Seppelt, R., Müller, F., Schröder, B., Volk, M.: Challenges of simulating complex environmental systems at the landscape scale: A controversial dialogue between two cups of espresso. Ecological Modelling **24**, 3481–3489 (06 2009). https://doi.org/10.1016/j.ecolmodel.2009.09.009
15. Wang, B., Chen, L., Yang, B.: Dm-nerf: 3d scene geometry decomposition and manipulation from 2d images (2023)
16. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
17. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. CoRR **abs/1801.03924** (2018), http://arxiv.org/abs/1801.03924
18. Zhang, X., Chen, Z., Wei, F., Tu, Z.: Uni-3d: A universal model for panoptic 3d scene reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9256–9266 (October 2023)