Multiple Multi-Modal AI for Semantic Annotations of 3D Spatial Data

Lee Kent[a], Hermenegildo Solheiro[b] and Keisuke Toyoda[c]
Toyoda Lab, University of Tokyo, Tokyo, Japan

Keywords: Object Detection, Scene Understanding, Semantic Annotation.

Abstract: 3D reconstruction of physical environments presents significant challenges, particularly when it comes to the semantic interpretation of these spaces, which often requires human input. This paper introduces a novel process that leverages multiple AI models trained on 2D images to automatically interpret and semantically annotate 3D spaces. Using a game engine as an intermediary, the process facilitates the integration of various 3D formats with 2D-trained AI models, enabling the capture and reprojection of semantic annotations back into the 3D space. A representative 3D scene is employed to evaluate the system's performance, achieving an object identification accuracy of 87% alongside successful semantic annotation. By offloading semantic annotation tasks to external 2D AI, this approach reduces the computational burden on edge devices, enabling dynamic updates to the system's internal knowledge base. This methodology enhances the scalability of spatial AI, providing a more comprehensive understanding of 3D reconstructed environments and improving the feasibility of real-time, AI-driven reasoning in spatial applications.

1 INTRODUCTION

A variety of sensors and techniques exist to capture and digitally represent physical environments. However, the process of accurately virtualizing the physical world involves navigating numerous challenging and often conflicting requirements. For instance, trade-offs must be considered between factors such as capture fidelity, coverage area, refresh rate, the types of physical sensors employed, and the utility of the resulting file format. The process of translating the physical world into virtual representations is commonly referred to as 3D reconstruction (Han et al., 2021; Sun et al., 2021).

While numerous methods for 3D reconstruction of physical spaces are available, some of which are detailed in Section 2, few automated approaches go further and integrate semantic data into these virtual models. This integration, which could significantly enhance the utility and understanding of reconstructed scenes, remains largely unrealized. For example, following the 3D reconstruction of an interior using LiDAR data, the individual objects within the geometry can be identified and labelled with additional properties such as materials, functional context or whether they are moveable. The semantic knowledge increases the model's utility as an identified chair can be indexed for searchability, extracted from the complete 3D reconstruction, customised, moved, hidden, or replicated, enabling advanced interaction and analyses.

The wide range of fundamentally different formats and purposes means that systems with general understanding of 3D spaces is a significant challenge, and each different format may require a completely different technique (Han et al., 2021). This is made more prominent when considering how AI models are trained and used. The data that an AI is trained with must be interoperable with the test data. For example, an AI trained only on point cloud data will only be able to interpret point cloud data.

Being able to work beyond the constraints of datatypes and sensors could prove invaluable when trying to reconstruct 3D space. Figure 1 presents the

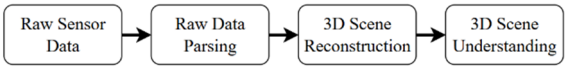


Figure 1: Towards 3D scene understanding.

[a] https://orcid.org/0000-0001-8546-547X
[b] https://orcid.org/0000-0002-0146-6684
[c] https://orcid.org/0009-0009-4434-1329

Kent, L., Solheiro, H. and Toyoda, K.
Multiple Multi-Modal AI for Semantic Annotations of 3D Spatial Data.
DOI: 10.5220/0013235300003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 1: GRAPP, HUCAPP and IVAPP*, pages 308-316
ISBN: 978-989-758-728-3; ISSN: 2184-4321

abstract steps required to achieve 3D scene understanding. This paper assumes that the 3D scene reconstruction is already complete and proposes a filetype agnostic approach to deriving semantic data.

Large strides have been made recently with regards to 2D image recognition in specific contexts, such as object identification and relative localisation using SLAM. In these areas AI has become quite proficient, both in terms of capability and speed (Gemini Team et al., 2024). This progress is enabled by the much larger pools of images available for training data, the equivalent of which does not exist in 3D formats. This advancement of 2D understanding of space can be leveraged in order to increase understanding and interaction with 3D spaces, and its value can be extended to domains such as to facilitate XR interactions (Sun et al., 2021), mechanical engineering (Kent et al., 2021), robotics (Batra et al., 2020; Weihs et al., 2020), and spatial AI (Hubert et al., 2021; Miyake et al., 2023).

This paper will describe a process to understand any 3D room reconstruction in any data format, employing several 2D image-based AI to identify and tag objects in a 3D reconstructed space with semantic information. The core contribution of this paper is the process of using 2D projections in captured 3D space to facilitate interpretation and reasoning of 3D spaces. The paper will present an implementation of the process for analysis and discussion.

This section has described the challenges with 3D scene understanding, and how these challenges are compounded by the need for many data formats for 3D reconstruction. Section 2 outlines current capability for scene understanding and defines the scope of this work. Section 3 details the proposed process, a complete implementation, and a testing scenario. Section 4 presents the findings and analyses the capability of the demonstrated implementation, along with discussion and iterations. This is followed by a discussion on limitations and generalisability in Section 5. The paper concludes with opportunities for future work in Section 6.

2 RELATED WORKS

In this section, approaches to 3D reconstruction and understanding will be described. Data representation will be considered and compared against the depth of understanding that can be achieved using them by AI, summarised in Figure 2. Typically, as the dataset becomes richer, the depth of possible understanding via AI decreases.

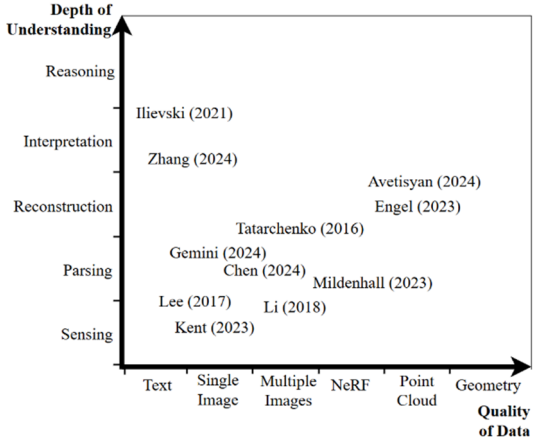


Figure 2: Mapping the related works quality of data used to the depth of understanding achieved.

Currently, text-based representations offer the best chance to enable implicit knowledge to be inferred by AI through structured knowledge graphs. Logic based tasks however, particularly with a temporal element, are still very challenging for AI (Jiang et al., 2023; Titus, 2024). Current methods consolidate a variety of knowledge modelling approaches in order to facilitate grounded contextual and temporal reasoning for AI (Ilievski et al., 2021). With methods to translate 3D space into semantic knowledge graphs, this could be utilised to increase an AI's faculty of spatial reasoning.

Several approaches look to parse 3D information from 2D images. For example, RoomNet (Lee et al., 2017), attempts to estimate 3D room layout from single 2D images by identifying key points within the space. Kent et al., (2023) propose identification of large structures and assemblies through smaller or standard components. This could be used to infer geometries of a room as part of a larger process chain. Depth can only be estimated in an image, so whilst identification of the objects is possible under the right conditions, there is no spatial or contextual interpretation or reasoning.

Li et al., (2018) uses silhouettes generated from 2D images to match to known object pools. By estimating depth from RBG images and combining multiple images, Tatarchenko et al., (2016), create explicit and complete point clouds of objects, although without explicit knowledge of what object is being created. NeuralRecon (Sun et al., 2021) extend this approach, and demonstrate real-time 3D reconstruction of 3D surfaces from a singular monocular video, analysed as a sequence of images. Depth in estimated in each image individually, leading to noisy outputs and redundant computation.

Work is also being done to understand 3D scenes where unordered or raw 3D data is available, such as

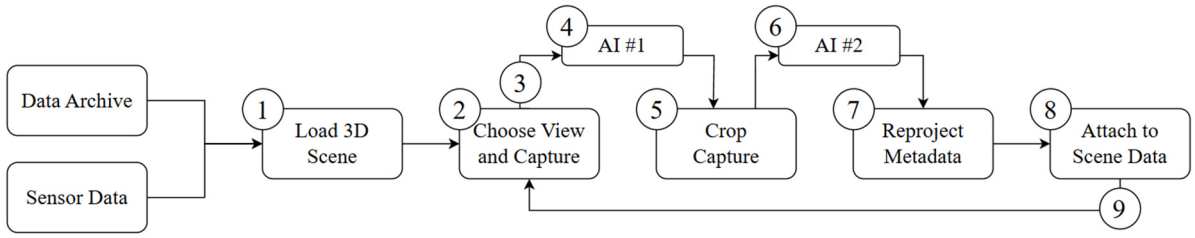


Figure 3: Process using multiple AI to develop scene understanding. The numbered steps are described in Table 1.

with point cloud captures. SceneScript (Avetisyan et al., 2025) uses LLM style next token prediction in order to describe point clouds as parametric instructions. This is a lightweight approach to scene definition, using high-level and interpretable representations of the space. Synthetic point cloud have been used as training data for an AI that can generate 3D layout estimations within the bounds of the parameters (Engel et al., 2023). This approach can build simple models without contextual knowledge, however manual extensions to achieve this has been proven theoretically possible.

Neural Radiance Fields (NeRFs) have recently gained significant attention (Mildenhall et al., 2022). NERFs can render novel views by training a deep neural network from a set of known camera poses. It has been followed up by Gaussian Splatting (Chen & Wang, 2024; Kerbl et al., 2023), which enables 3D reconstruction from images, offering improvements in computation speed and data size. Whilst these methods provide dynamic and fast 3D reconstruction, they are incompatible with current 3D understanding methods.

When these myriad approaches are mapped, see Figure 2, there is a disconnect between having high quality data through 3D reconstruction, and the depth of understanding. By reducing high quality data to data that is compatible with AI models, the output data can be reprojected into the reconstructed space.

In summary, 2D images have shown potential as inputs for semantic understanding, but too much data is required for a general AI capable of 3D scene interpretation. Point clouds can be used to capture a scene but are data heavy and lack semantic understanding methods. NeRFs and Gaussian Splatting have showed promise as a means to represent 3D scenes (Chen & Wang, 2024), but the generated views are assumed, and also with little semantic understanding opportunities. This paper presents a process that combines the speed and opportunities for semantic understanding by representing a 3D space as a series of 2D projections.

3 PROPOSED PROCESS

The process described in this paper uses a game engine and a series of AI to parse and understand a 3D environment. The system will emulate egocentric exploration of a 3D virtual scene, attempting to parse and interpret the initially unidentified objects. Figure 3 shows the process being proposed, and Table 1 details each of the process steps.

This process shows two AI being used (Gemini Team et al., 2023, 2024), but this can be extended, and other AI can be substituted if required. The important distinction is the use of parallel captures being processed serially by AI, cropped, and reprojected back into the 3D scene to develop spatial understanding about the 3D scene. A prototype system has been developed to explore how this cropping and reprojection can be used in a 3D reconstructed scene.

3.1 Prototype System

Two state of the art AI (Gemini Team et al., 2023, 2024), are utilised to infer details about the objects in a 3D scene. These AI models are trained on 2D data, allowing them to operate effectively with 2D images. To leverage their capabilities, we will capture a series of egocentric 2D images of the 3D scene.

Following evidence that cropping images to the subject can increase the success and capability of zero shot inference for Multimodal Large Language Models (MLLMs) (Ilievski et al., 2021; Zhang et al., 2024), the first AI specifically identifies objects within the capture and provides bounding boxes with confidence values around any identified objects.

The initial capture is cropped to each bounding box to ask more precise questions to a second AI, which is able to infer more qualitative information such as colour, function, and context. The second AI is multimodal, accepting images and a text prompt for specific information. This AI is unable to describe the object location within the capture so to reproject the data, the object must be central in the capture.

From Figure 4, examples of bounding boxes leading to cropped images can be seen. These cropped images are then sent to the second AI. The text prompt uses the English order of adjectives list (Cambridge Dictionary, 2024) to populate semantic data and is as follows:

"What do you see in this image? Be as specific as possible. Return the results as a csv file in the format 'object name, opinion, size, physical quality, shape, age, colour, origin, material, type, purpose'."

The use of a second AI allows for scene understanding, supplementing the 2D capture with semantic data and context. The response is then reprojected back into 3D space and attached to the 3D geometries. Table 1 gives a more detailed overview of each step of the process and how it is implemented in the prototype system.

3.2 Testing Scenario

A test scenario will be used to evaluate the viability of using several multi-modal AI to generate data and semantic data about a 3D scene. As this paper is not concerned with the method for 3D reconstruction, an existing scene with known geometries and collisions for all objects will be used. This will ensure that the reprojected response can be analysed for correctness. The scene is a simple room with 68 internal and identifiable objects. Exterior boundaries, such as walls, floors, and ceilings are not identifiable by this process. There are areas of cluttered objects, such as items scattered on a table and books on a bookshelf.

Figure 4: Screenshot and resulting bounding boxes.

The process will run in batches of 75 cycles until the results stabilise, to explore how the scene understanding evolves over time. To understand the capability of the scene identification process, a trial run of 225 captures will be presented.

Table 1: Description of the complete process.

#	Step	Description
1	Load 3D Scene	Any prior 3D reconstruction data is loaded into the scene. In this prototype, the 3D scene has geometry and collision data.
2	Choose 2D capture position	A random position within the 3D space is selected, emulating an egocentric view.
3	Take 2D Screenshot	From the chosen 3D position, a 2D screenshot is taken. The image as well as the camera parameters and transform are both archived and sent to AI #1.
4	Send to AI #1 (Vision)	AI #1 identifies and segments objects. The response is a series of bounding boxes around objects see (Figure 4).
5	Crop 2D Capture	The screenshot in Step 3 is cropped for each bounding box. Each cropped image is sent to AI #2 with a text prompt.
6	Send to AI #2 (Gemini)	AI #2 can infer direct information about a single object, such as name, size, materials etc. As the image is cropped, there *should* be a single subject within each image.
7	Reprojection	The response data is parsed. A ray cast replicating the initial 2D capture is used to reproject the parsed data back into the 3D scene.
8	Attach to scene	The parsed data is then attached as metadata to any hit objects. The data attachment method depends on the 3D scene filetype.
9	Go to Step 2	Repeat until satisfied.

4 RESULTS

This section presents the outcomes of a pilot implementation of the process. In Section 4.1, the results are manually compared against known correct values to evaluate the accuracy of object identification. Section 4.2 explores the system's ability to measure self-confidence in its results without human intervention. This is followed by Section 4.3, which introduces an extension to enhance the identification capabilities of the overall process.

Finally, Section 4.4 evaluates the accuracy and significance of the generated semantic data.

4.1 Identification Accuracy

Table 2 summarises the output of the test scenario as the scene identification progressed. The first column specifies the number of captures sent to AI #1. The second column specifies the number of cropped images that are then send to AI #2. Column three is the number of identified objects by AI #2. Column four is the number of discarded identifications. The data from AI #2 is reprojected back into the 3D scene, and if this ray does not hit viable geometry (e.g. a wall, the floor), then it is discarded. Column 5 show the number of successful object identifications.

Table 2: Number of captures sent to AI #1, cropped images sent to AI #2, identifications, discarded results and successful identifications.

Captures	Crops	IDs	Discards	IDs
75	231	273	53	220
150	494	464	98	366
225	743	676	121	601

Figure 5 show the progression of object identification. By increasing the number of scenes captures, there is an increase in number of identified objects. However, the objects that are challenging for the AI to identify remain unidentified, even as the number of captures increases. Between 150 and 225 cycles, only one more object was able to be identified.

The results in Figure 5 are manually checked for accuracy. After 225 cycles; 28 of the 68 were never identified, in that no data was reprojected to them. This could be attributed to multiple factors and will be discussed in Section 4.3.

4.2 Self-Confidence of Identification

Something unaccounted for in the colour maps in Figure 5 is the self-confidence of the found values. Self-confidence is the level of certainty in the identification. This can manifest in multiple ways. A chair that was correctly identified 23 times out of 23 appears in Figure 5 has the same 100% self-confidence value as a book that was found once and identified correctly once. This self-confidence value needs to be measured, as the process should run independently. Following the 225 cycles, three factors limiting self-confidence have been identified. These are:

- Name synonyms (Section 4.2.1)
- Many items in cropped images (Section 4.2.2)
- Unidentifiable objects (Section 4.2.3)

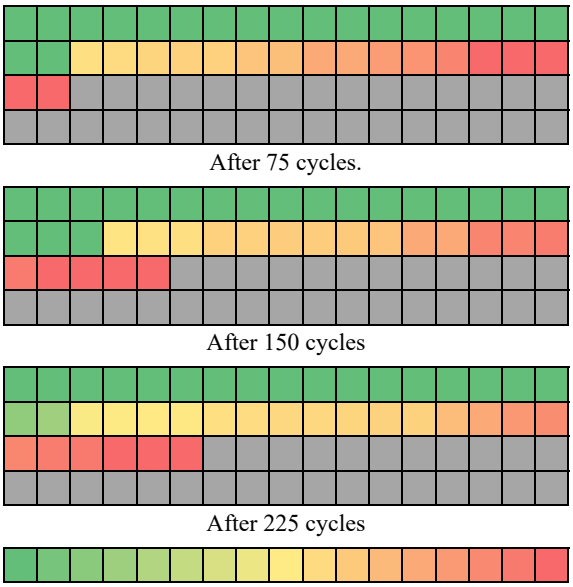


Figure 5: Progression of correct identification of 68 objects in a 3D scene. Figure best viewed in colour.

4.2.1 Name Synonyms

As the results from AI #2 does not account for synonyms of objects names, an object may be correctly identified but not align with previous identifications. For example, after 225 captures, the 'sofa' in the room was identified correctly 45/57 times, but using several different terms, see Table 3. The most accurate ID, Chaise Lounge, is arguably the most correct, but the other 'correct' identifications will still reduce the overall confidence.

This discrepancy in not-false positives may be overcome by calculating semantic distances between words, and using that value to determine confidence using knowledge graph networks such as WordNet (Princeton University, 2010) or ConceptNet (Speer et al., 2017).

Table 3: Differently correct identifications of the same object.

Identified Name	Count
Chaise Lounge	27
Sofa	14
Couch	3
Red Velvet Chair	1

4.2.2 Many Items in Cropped Images

Some objects were found to have a wide range of conflicting identifications. This was particularly

prominent when the object was a container or a surface. In these instances, the object was not the sole subject of the cropped image, a known weakness (Ilievski et al., 2021; Zhang et al., 2024). As the data is attached to the scene using the centre of the cropped image, this leads to incorrect attachments to containers and surfaces.

In the test scene there are three tables. The self-confidence was 18-21%, with a combined total of 14/72 correct identifications. From Table 4, it is clear that the objects on and around the tables are being associated with the tables, significantly reducing the overall confidence.

Table 4: Identifications and their relation to the attached object.

	Named object in identification is:		
	Correct	On Table	Next to Table
Table 1	3	1	10
Table 2	7	23	8
Table 3	4	9	7

This error could be overcome by adding a decision point at Step 7 in Figure 3. In the case where multiple items are identified in the cropped image; the cropped image could be resubmitted to AI #1 to further reduce into sub-images. This would help remove the objects on and around the table from the tables ID data.

On inspection of all 14 correct table identifications in Table 4, it was never the sole subject of a cropped response. This means that adding cropped images with multiple IDs back into the pool would remove any chance of identifying a container or surface. This would require the use of a second prompt targeted at surfaces and containers. Alternatively, objects with high self-confidence could be removed from the scene.

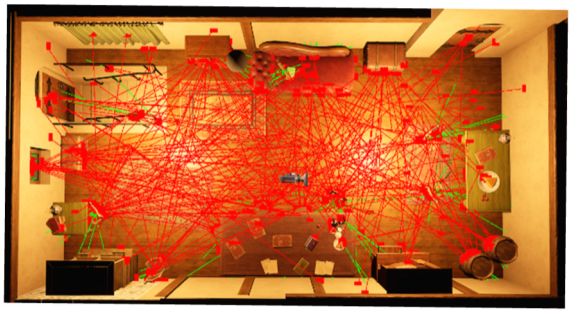


Figure 6: Top-down view of the scene used in the test scenario. Each represents a reprojection of semantic data into 3D space from a capture position. Best viewed in colour.

4.2.3 Unidentifiable Objects

There are some objects that AI cannot or cannot consistently identify. This can be for myriad reasons, such as not being in a capture, poor, missing, or noisy input data, occlusion, or the object is simply not in the AIs capability to identify.

After 250 captures, the remaining 28 unidentified objects are all small objects that occlude each other, for example stacks of books or plates under cups. Manually selecting ideal screenshots have shown that under the right conditions, the objects can all still be identified correctly. The current capture selection algorithm selects random positions and rotations within the space. Replacing this randomness with a more structured approach is expected to improve the overall effectiveness of the process.

4.3 Informed Capture Selection

Until now, the selection of views for captures has been random. Low self-confidence can be added as weighted selection criteria to the capture selection step. The self-confidence of each object is calculated, and the less self-agreement there is in within the data attached to an object, the more likely it should be within the next captures frame. This was achieved by finding the closest object in the lowest 5% of self-confidence and linearly interpolating the camera transform towards the object.

The extension was run for a further 75 captures. This led to the identification of a further 25% of the objects, with 87% of the objects in the room now being identified with 59% having self-confidence in the ID over 50%. Figure 7 shows the small effective difference between correct identification percentage and self-confidence in identification after 300 cycles.

The objects with zero correct identifications after 300 cycles are a book, a carpet, 2 plates, 3 flowers, 3 pieces of paper and a vase. Most of these can be traced down to occlusion, something that cannot be overcome using optical sensors for 3D reconstruction. The books spine is difficult to differentiate against the other books. The flower stems are thin, so it is unsurprising that the raycasts consistently ‘missed’ the location, hitting the out of bounds exterior walls. The paper and vase not being identified are outliers, without an apparent reason the AI could not identify them. Possibly, they were unfortunate and not included in any capture, or the AI is simply unable to identify them, even under perfect conditions. The extension to the informed capture selection ensured that the objects and areas with less self-confidence

were given focus and this had a clear impact on the results, improving the viability of the approach.

4.4 Capturing Semantic Data

The final part of this section concerns the capture of additional semantic data, providing additional meaning and context to the identified geometry. More than being able to identify the information, it is valuable to capture tacit information. Recent advances in 2D AI have begun to facilitate this (Gemini Team et al., 2024).

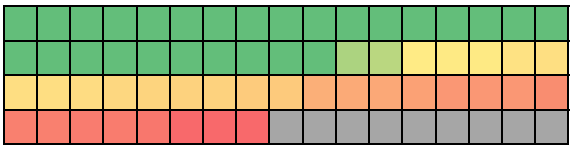
For this proof of concept, the English order of adjectives was used to get additional information about the seen objects (Cambridge Dictionary, 2024). The *size* of the object was generally always in self-agreement, despite only being able to assume the scale. A giant apple would likely still be classified as small. Interestingly, the 2D AI occasionally tried to give specific dimensions, which unsurprisingly were nowhere close to accurate.

Physical quality typically described the surface of the object, or for around half of instances simply said 'solid.' *Shape, Age* and *Origin* were either 'unknown' or incorrect guesses and were included for completeness.

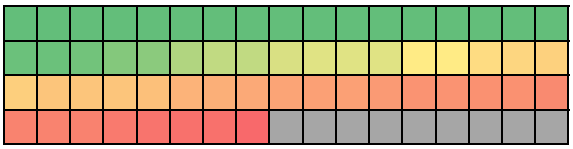
The *Material* of each object was identified, but in vague categories, such as 'Wood' or 'Ceramic,' without specifics such as grain or deterioration. When paired with colour or other semantic data, realistic or predefined textures could be applied as part of a 3D reconstruction process, using textures and materials applied from 3D libraries.

Type and *Purpose* provide means to categorise the objects further. Automatic identification is helpful, but being able to further contextualise the objects can provide more rich value, for example in the push for spatially aware AI. For instance, correct barrels identifications also supplied a range of additional semantic information such as: 'container, cask, storage, to store liquids, for transport.'

Spatial arrangements were also able to be identified with an addition to the prompt: 'positional relationships between the objects. Adjacencies and relative positions can be helpful when providing instructions on how to achieve tasks, for example to direct towards a specific location 'on a table' or 'opposite the doorway.' The spatial arrangements must be captured before the cropping step, but then processed and attached to the data after the cropping step. In future work, the combinatorial use of this semantic data will be used as input for Spatial AI with faculty for 3D spatial reasoning. This will be described in Section 6.



Successful identification of 68 objects after 300 process cycles.



Self-confidence in identification of 68 objects after 300 process cycles.

Green (Left) indicates 100% accurate identification, red, 0%, and grey boxes were never identified and had no attached data.

Figure 7: Comparing the difference between successful identification and self-confidence in that identification after 300 cycles.

5 LIMITATIONS

In this section, we discuss the generalisability and the limitations of the process and the study.

One of the key challenges for scene identification is the considerable number of representations and filetypes for 3D spaces. The process in Figure 3 is datatype agnostic, assuming the datatype can be loaded into a game engine. For this testing scenario, a scene with separate geometries for each object was utilised.

A common datatype for this kind of scene is point clouds, due to the range of sensors available. For example, many VR headsets have on board point cloud sensors that are used for localisation. The proposed process in this paper could be run on these point clouds dynamically.

The attachment of data to the 3D file will remain datatype dependant. For example, instead of a raycast, a mask will be required for a point cloud, and the reprojection of semantic data will need to be associated with all points within the mask reprojection. This could significantly impact storage requirements if not managed appropriately.

Only one room was tested for this pilot, to explore the feasibility of the approach, focusing on depth of analysis and challenges. By running the process on many rooms in different configurations and from various origins, the potential of this approach can be further scrutinised and improved.

The AIs used is also limiting factor. They can only identify objects that they are trained to identify and that have sufficient representation in the training data. The process should be AI agnostic, using web interfaces to connect to two AIs with differing capabilities. There are alternative AI candidates that could be used (Kirillov et al., 2023; Redmon et al., 2016). The novelty of this papers is the dynamic cropping procedure and the reprojection back into the scene.

Finally, offloading the processing to external servers, which involves streaming 3D scenes and the people within them, conflicts with privacy protections. Consequently, many HMDs by default do not allow capture or recording of passthrough sensors. Whilst enabling this capability unlocks many opportunities, it requires strict adherence to privacy regulations which are foundational to protect individual rights, and maintaining public trust must be a priority in all applications.

6 CONCLUSIONS

Multiple multi-modal AI were used that had different inputs and differing capabilities. Both are only capable with 2D images; however, their combinatorial use enabled the identification of objects in 3D space. After 300 cycles 87% of objects were correctly identified, albeit because of noun synonyms, only 59% having self-confidence over 50%. Semantic data was also captured, providing a range of descriptors, object use and positional relationships to other objects.

A Game Engine acted as the intermediary system, with the ability to load many 3D formats and to interface with cloud-based AI systems. The addition of weighted capture selection towards objects with lower self-confidence improved the process.

The semantic data was reprojected back into 3D space and attached to the objects in the scene. Objects with predefined geometries and collisions were used in this study, but the process could be extended to other datatypes, for example, using masks and point clouds. The value of the process is the reprojection and attachment of the 2D AI findings back into 3D space, utilising 2D AI capabilities in 3D contexts.

6.1 Future Work

Through this paper, several opportunities for future work have been identified. The process can be improved with an additional step to ensure that exteriors, surfaces, and containers are appropriately identified. The outcomes may also be improved by using semantic closeness of the AI responses, to have more representative and accurate self-confidence measures.

Other datatypes should also be considered, such as NeRFs and point cloud clouds. This approach is suitable for any 3D datatype interoperable with games engines but has only been implemented using a scene with static geometries. The tagging process will depend on the datatype.

Substantial work is being conducted in areas such as multi-modal large language models, commonsense reasoning and spatial AI, that look to facilitate and enhance everyday tasks. This process of dynamically interpreting 3D scenes and translating to human readable semantic information can serve as spatial inputs to these advanced and complex models.

ACKNOWLEDGEMENTS

The authors would like to thank the Common Ground Living Lab members for their support.

REFERENCES

Avetisyan, A., Xie, C., Howard-Jenkins, H., Yang, T.-Y., Aroudj, S., Patra, S., Zhang, F., Frost, D., Holland, L., Orme, C., Engel, J., Miller, E., Newcombe, R., & Balntas, V. (2025). SceneScript: Reconstructing Scenes with an Autoregressive Structured Language Model. *Computer Vision – ECCV 2024*, *15119*, 247–263. https://doi.org/10.1007/978-3-031-73030-6_14

Batra, D., Chang, A. X., Chernova, S., Davison, A. J., Deng, J., Koltun, V., Levine, S., Malik, J., Mordatch, I., Mottaghi, R., Savva, M., & Su, H. (2020). *Rearrangement: A Challenge for Embodied AI* (arXiv:2011.01975). arXiv. http://arxiv.org/abs/2011.01975

Cambridge Dictionary. (2024). *Order of adjectives*. https://web.archive.org/web/20240404112407/https://dictionary.cambridge.org/grammar/british-grammar/adjectives-order

Chen, G., & Wang, W. (2024). *A Survey on 3D Gaussian Splatting* (arXiv:2401.03890). arXiv. http://arxiv.org/abs/2401.03890

Engel, J., Somasundaram, K., Goesele, M., Sun, A., Gamino, A., Turner, A., Talattof, A., Yuan, A., Souti, B., Meredith, B., Peng, C., Sweeney, C., Wilson, C., Barnes, D., DeTone, D., Caruso, D., Valleroy, D., Ginjupalli, D., Frost, D., … Newcombe, R. (2023). *Project Aria: A New Tool for Egocentric Multi-Modal AI Research*. https://doi.org/10.48550/ARXIV.2308.13561

Gemini Team, Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., … Vinyals, O. (2023). *Gemini: A Family of Highly Capable Multimodal Models*. https://doi.org/10.48550/ARXIV.2312.11805

Gemini Team, Reid, M., Savinov, N., Teplyashin, D., Dmitry, Lepikhin, Lillicrap, T., Alayrac, J., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., Antonoglou, I., Anil, R., Borgeaud, S., Dai, A., Millican, K., Dyer, E., Glaese, M., … Vinyals, O. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*. https://doi.org/10.48550/ARXIV.2403.05530

Han, X.-F., Laga, H., & Bennamoun, M. (2021). Image-based 3D Object Reconstruction: State-of-the-Art and Trends in the Deep Learning Era. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(5), 1578–1604. https://doi.org/10.1109/TPAMI.2019.2954885

Hubert, T., Schrittwieser, J., Antonoglou, I., Barekatain, M., Schmitt, S., & Silver, D. (2021). Learning and Planning in Complex Action Spaces. *International Conference on Machine Learning*. https://doi.org/10.48550/ARXIV.2104.06303

Ilievski, F., Szekely, P., & Zhang, B. (2021). CSKG: The CommonSense Knowledge Graph. In R. Verborgh, K. Hose, H. Paulheim, P.-A. Champin, M. Maleshkova, O. Corcho, P. Ristoski, & M. Alam (Eds.), *The Semantic Web* (Vol. 12731, pp. 680–696). Springer International Publishing. https://doi.org/10.1007/978-3-030-77385-4_41

Jiang, Y., Ilievski, F., Ma, K., & Sourati, Z. (2023). *BRAINTEASER: Lateral Thinking Puzzles for Large Language Models* (arXiv:2310.05057). arXiv. http://arxiv.org/abs/2310.05057

Kent, L., Snider, C., Gopsill, J., Goudswaard, M., Kukreja, A., & Hick, B. (2023). A Hierarchical Machine Learning Workflow for Object Detection of Engineering Components. *Proceedings of the Design Society*, *3*, 201–210. https://doi.org/10.1017/pds.2023.21

Kent, L., Snider, C., Gopsill, J., & Hicks, B. (2021). Mixed reality in design prototyping: A systematic review. *Design Studies*, *77*, 101046. https://doi.org/10.1016/j.destud.2021.101046

Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). *3D Gaussian Splatting for Real-Time Radiance Field Rendering*. https://doi.org/10.48550/ARXIV.2308.04079

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). *Segment Anything* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2304.02643

Lee, C.-Y., Badrinarayanan, V., Malisiewicz, T., & Rabinovich, A. (2017). *RoomNet: End-to-End Room Layout Estimation*. https://doi.org/10.48550/ARXIV.1703.06241

Li, K., Garg, R., Cai, M., & Reid, I. (2018). *Single-view Object Shape Reconstruction Using Deep Shape Prior and Silhouette*. https://doi.org/10.48550/ARXIV.1811.11921

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2022). NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, *65*(1), 99–106. https://doi.org/10.1145/3503250

Miyake, Y., Toyoda, K., Takashi, K., Hyodo, A., & Seiki, M. (2023). Proposal for the Implementation of Spatial Common Ground and Spatial AI using the SSCP (Spatial Simulation-based Cyber-Physical) Model. *2023 IEEE International Smart Cities Conference (ISC2)*, 1–7. https://doi.org/10.1109/ISC257844.2023.10293487

Princeton University. (2010). *About WordNet*.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788. https://doi.org/10.1109/CVPR.2016.91

Speer, R., Chin, J., & Havasi, C. (2017). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). https://doi.org/10.1609/aaai.v31i1.11164

Sun, J., Xie, Y., Chen, L., Zhou, X., & Bao, H. (2021). *NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video*. https://doi.org/10.48550/ARXIV.2104.00681

Tatarchenko, M., Dosovitskiy, A., & Brox, T. (2016). Multi-view 3D Models from Single Images with a Convolutional Network. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (Vol. 9911, pp. 322–337). Springer International Publishing. https://doi.org/10.1007/978-3-319-46478-7_20

Titus, L. M. (2024). Does ChatGPT have semantic understanding? A problem with the statistics-of-occurrence strategy. *Cognitive Systems Research*, *83*, 101174. https://doi.org/10.1016/j.cogsys.2023.101174

Weihs, L., Salvador, J., Kotar, K., Jain, U., Zeng, K.-H., Mottaghi, R., & Kembhavi, A. (2020). *AllenAct: A Framework for Embodied AI Research* (arXiv:2008.12760). arXiv. http://arxiv.org/abs/2008.12760

Zhang, H., Du, W., Shan, J., Zhou, Q., Du, Y., Tenenbaum, J. B., Shu, T., & Gan, C. (2024). *Building Cooperative Embodied Agents Modularly with Large Language Models* (arXiv:2307.02485). arXiv. http://arxiv.org/abs/2307.02485