Simulating human visual imagination:

Scaling and placement of objects in procedural generation of 3D scenes

Taeko Jessica Bourque

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the
requirements for the degree of

Master

in

Cognitive Science

Department of Cognitive Science

Carleton University

Ottawa, Ontario

© 2021

Taeko Jessica Bourque

Abstract

Although there is already quite a bit of research on the uses of imagination, there is a gap involving how exactly mental images are created (Davies et al., 2011), and more specifically, how humans use their imagination to create scenes. This thesis describes the placement and scaling of objects in the creation of scenes in the visual imagination. To examine this topic, I created 30 images in Unity using scene descriptions generated by SOILIE 3D and 30 images that applied additional constraints based on an improved model of placement and scaling of objects. Contrasting these pairs illustrates the key aspects of human scene creation and demonstrates that my model is more realistic. While the previous model was too theoretical and failed to take many aspects of human cognition into account, applying research in neuroscience and human scene understanding has allowed for an improved model that is more consistent with a human perspective of scenes.

Acknowledgments

This work would not have been possible without the guidance and support of such a wonderful group of people that I have the good fortune of being surrounded by. First and foremost, I would like to thank my supervisor, Dr. Jim Davies. From being the voice of reason during my panicked ramblings, to answering even the most basic of questions with utmost consideration, your tireless support has guided me through these tumultuous times. I have been benefiting from your advice since our first meeting, when you encouraged me to pursue a master's degree and work with you. Were it not for you, I would still be trudging through a second degree with the hopes of one day being where I am now.

These past years I have had the opportunity to work alongside and learn from members of the academic community within Carleton University. I would like to thank Dr. Ida Toivonen for taking the time to be on my committee and for always asking the hard questions. While I undoubtedly still make mistakes when choosing my words, I am guided by your voice, questioning whether I am truly saying what I mean to say. Thanks to Dr. Robert West, for also agreeing to be on my committee and for your helpful suggestions on clarifying my manuscript. Thank you Dr. Craig Leth-Steensen for agreeing to be the Internal Examiner on my committee and keeping me on my toes with your questions. I will never think of Spongebob the same. Thanks to Dr. Alex Kelly for your comments and organizing my defense as Chair of my committee.

Thanks to SOIL members for their support and lively discussions during lab meetings. Special thanks to Nico Turcas for our 'SOILIE support group' chats. I would also like to thank Mike Cichonski for sharing the original SOILIE 3D code and helping me troubleshoot it. Thanks to the friends I've made along the way. To Lara Russo, Mujahid 'Gboye' Sanni, and Ashley Promislow, whether it was keeping each other apprised of deadlines, practicing our presentations, or just checking in on each other via Zoom, it was comforting to work through this together.

I could not have accomplished this without the love and support of my family. A big thank you to my parents, Laura Mitoma and Sylvain Bourque. Kaasan, thank you for your unwavering support (in all that I do, but especially during these crazy times). You've always inspired me to put forth my best effort and never give up. Thank you for being patient with me, whether I was frustrated with the code or overly excited about minute changes, and for brainstorming ideas with me. Merci papa pour les multiples discussions et tes suggestions concernant la compréhension de passages dans ma thèse et de mes graphiques. Ça fait longtemps que tu m'as montré le poème de Nicolas Boileau, mais "ce qui se conçoit bien s'énonce clairement" m'est toujours resté dans la tête. Thank you to my husband, Corey Smith, for being my rock and for always believing in me and picking me up when I had my doubts. I am immensely grateful for your patience in reading my multitude of drafts and workshopping ideas at all hours of the day (and night!). Even if you didn't have all the answers, you always knew exactly how to help. To my sister, Katie Annett, thank you for being there for me, proofreading passages, and helping me shorten my abstract!

This work is dedicated to my brother, Alex Bourque, without whom none of this would be possible. There's no one else

I'd rather have by my side when struggling with VMs, buggy code, and the dreaded OOM Killer. I'm still amazed that you built me a Linux computer solely for running this code! You helped me pass what seemed like insurmountable obstacles and I will be forever grateful. Thank you Baba. VER!

Table of Contents

Abstract	ii
Acknowledgments	iii
List of Tables	vii
List of Figures	viii
List of Equations	ix
List of Appendices	x
1 Introduction	1
1.1 Imagination	1
1.2 Mental imagery	1
1.3 Imagination without the generation of a mental image	3
2 Previous model	4
2.1 Science of Imagination Laboratory’s Imagination Engine: SOILIE	5
2.2 SUN3D database as input for SOILIE 3D	5
2.3 Step 1: collection of data	6
2.4 Step 2: preparation of data	7
2.5 Step 3: placement of objects and rendering a final 3D scene	8
2.6 Output	10
2.7 Notable Issues	11
2.7.1 Data limitations: Small total number of objects	12
2.7.2 Object parts versus wholes	12
2.7.3 Object placement	13
2.7.4 Object scaling	14
3 Improved model	14
3.1 Improving SOILIE 3D output	14
3.2 Before scenes	15
3.3 After scenes	16

3.3.1	Object placement	17
3.3.2	Object scaling	25
4	Results	27
4.1	Evaluating the output	27
5	Discussion and conclusion	29
5.1	General discussion	30
5.1.1	AI vs the human mind	30
5.2	Limitations	34
5.3	Related work	37
5.4	Future Work	39
5.5	Conclusion	39
6	References	41
7	Appendix A: 30 pairs of before and after images of scenes	44

List of Tables

1	Distances and angles	8
2	Centroid locations	8
3	Sample objects and respective size categories	26
4	Size Categories and Scaling Factor	27

List of Figures

1	Sample CSV from one scene in the SUN3D database	6
2	(2013).[Frame from one scene in the SUN3D database]. SUN3D Database. http://sun3d.cs.princeton.edu/ . . .	6
3	Sample CSV created from one frame within one scene in the SUN3D database	7
4	Sample output of triplets	7
5	Sample output of size	7
6	SOILIE runtime menu	10
7	Sample of working combinations	10
8	Sample image output of a generated scene	11
9	Sample CSV output of a generated scene	11
10	Problematic output from SOILIE 3D	13
11	Sample CSV files used for creation of following scenes	15
12	Before image of Scene 1	16
13	After image of Scene 1	17
14	Object binding areas	20
15	Chair below table	20
16	Decision flowchart for intersecting objects	21
17	Problematic before scene with objects placed unnaturally far apart	22
18	Improved scene with objects placed closer together	23
19	Problematic scene with blinds placed in center of a room	23
20	Improved scene with blinds placed along the bounding box of the room	24
21	Improved but still problematic window and blinds placement	25
22	Estimated object dimensions in inches	26
23	Example of a believable scene	28
24	Moderately improved scenes	29
25	Before and after images of scenes	44
25	Before and after images of scenes	45
25	Before and after images of scenes	46

List of Equations

1	calculating the x-coordinate of object C	9
2	calculating the y-coordinate of object C	9
3	calculating the x-coordinate of observer (object O)	9
4	calculating the z-coordinate of observer (object O)	9
5	calculating the z-coordinate of observer (object O)	9

Appendices

Appendix A: 30 pairs of before and after images of scenes	46
---	----

1 Introduction

Humans use their imagination daily, across many fields, and it plays an important role in the development of the individual as well as that of society (Archambault & Venet, 2007). There are many cognitive abilities for which imagination plays an important role, such as problem solving, planning, mental time travel, hypothetical thinking, counterfactual thinking, and theory of mind (Davies, Atance, & Martin-Ordas, 2011).

The goal of the present work was to simulate human imagination by procedurally generating plausible 3D scenes to illustrate the underlying aspects of visual mental imagery used in scene creation. This paper is divided into five sections: a brief overview of imagination and visual mental imagery, a description and analysis of the lab's previous model of imagination (SOILIE 3D), a description of the methods I used to improve this model and the theories supporting these changes, the resulting comparison of images produced, and a discussion of how this approach approximates human imagination.

1.1 Imagination

Imagination can be seen as the creation of ideas in our minds that are made up of our memories, beliefs, and other ideas (Davies, 2019). While imagination is often treated as synonymous with creativity ("What a wild imagination she has!"), that is only one aspect of it. Imagination has been defined as being comprised of two parts: that which reproduces reality in the form of images (visual imagery), or that which associates images to create something new (creativity) (Archambault & Venet, 2007; Davies et al., 2011; Lalande, 1991). For the purpose of this thesis, imagination will not be used in the sense of 'creativity' but rather in the form of mental imagery.

1.2 Mental imagery

Aristotle considered an image (stored in memory and substituting for the object) to be indispensable because it allows us to think in real time that which is timeless; imagination allows thought removed from the constraints of time and space (Archambault & Venet, 2007). When you think back on your 9th birthday, you are using your imagination to recreate a moment in space and time that no longer exists (Davies, 2019).

Visual imagery is also known as "seeing with the mind's eye" and this mental simulation is important for both remembering the past and thinking of the future (Dawes, Keogh, Andrillon, & Pearson, 2020; Keogh, Bergmann, & Pearson, 2020). There are many other situations in which visual imagery is used: in mental practice, by training or recovering athletes, for therapy treating anxiety or PTSD, when trying to recall the face of a neighbour or the layout of a house, etc. You can consciously choose to imagine a big fluffy dog without having one to look at. You can also imagine something without consciously choosing to do so; imagery can come automatically in some situations, such as when reading a book, or after experiencing

trauma (Davies, 2019). Watzlawic (1988) states that our representation of reality is influenced by what we are; our mental representations are unique because our representations of the world and everything in it are dependent on our experiences. For example, my mental representation of "dog", "flower" or "winter" will be different than that of a Belgian florist or of a Brisbanian's.

Semantic declarative memories (e.g., knowing that a roof goes on top of a house) are considered to be represented as sentence-like structures known in cognitive science as propositions. They are made up of symbols (representing things such as shape, distance, colour, etc.) that are arranged in a meaningful structure that shows how they are related. This is considered to be at the core of all of our long-term memories (Davies, 2019).

It is now mostly agreed upon that our memories (including visuospatial) are represented descriptively (Davies et al., 2011). Further evidence that memories are not entirely depictive (but rather composed of symbolic structures) is that we can retrieve memories of specific scenes based solely on a word mentioned (Davies et al., 2011).

Visual mental imagery has been a rather controversial subject in cognitive science. Some of the debated issues are whether images differ fundamentally from verbal thoughts, whether they share underlying mechanisms with visual perception, and whether the information in these images is represented in a spatial (map-like) format (Farah, 1989; Pylyshyn, 2002).

There has been research on the neural systems underlying imagery that sheds light on these controversies as well as on the cerebral localization of imagery processes (Farah, 1989). Psychologists and neuroscientists have shown that visual imagery can be depictive in nature (Pearson & Kosslyn, 2015) and that visual imagery is closely related to many cognitive functions (e.g., visual memory, spatial navigation, language comprehension, making moral decisions) (Keogh & Pearson, 2018). It has even been shown that using visual imagery (imaginal exposure, imaginal rescripting) during cognitive behavioural therapy is more effective than just verbal processing (Keogh & Pearson, 2018).

It is now agreed that people can experience visual images, that propositional representations can be used in cognition, and that a depiction (picture or mental image), if it exists, must be interpreted in a specific way in order for it to have meaning, which requires a propositional component (Kosslyn, 1994). Now that we have evidence from neuroscience showing what images are (and where they are processed) (Kosslyn, 1994; Pylyshyn, 2002) many people believe the imagery debate to be resolved. This resolution to the mental imagery debate is in itself debatable, especially considering that not everyone agrees with the exact subject of the debate. Pylyshyn (2002) argues that the claims regarding the resolution of the mental imagery debate (regardless of the exact subject) haven't been sufficiently explored enough to arrive at a solution.

Although it is generally agreed that many people experience mental imagery, there is still a debate regarding the existence of a depictive representation. Pylyshyn (2002) remains a main opponent of a depictive representation of mental imagery and believes that the processes and representations underlying imagery are the same as other forms of reasoning.

To assess the maximal degree of shared neural processing in visual mental imagery and visual perception, Ganis, Thomp-

son, and Kosslyn, (2004) observed participants with fMRI after they had either visualized or seen faint drawings of simple objects and while they judged specific aspects of the drawings. Their results revealed that visual imagery and visual perception draw mostly on the same neural machinery, although the spatial overlap was incomplete and irregular. They discovered that the overlap was much more pronounced in the frontal and parietal regions than in the temporal and occipital regions. This suggests that cognitive control processes function comparably in both imagery and perception, while some sensory processes are engaged differently by visual imagery and perception (Ganis et al., 2004).

When viewed from an information processing point of view, it is difficult to separate the cognitive processes belonging to rational thought and those belonging to imagination (Archambault & Venet, 2007). We have the experience of “seeing” in both perception and imagination, with the main distinction being that imagination is a less-vivid recreation based on memories and not something you are currently perceiving (Davies, 2019).

1.3 Imagination without the generation of a mental image

It is important to note that human imagination does not always include the generation of a mental image; imagination is not reliant on imagery. Not only is it possible to have non-sensory imagination, but our imaginings are usually a combination of sensory (mental imagery) and non-sensory (conceptual imagination) information (Davies, 2019). Our imagination is composed of belief-like entities which could be of a semantic and propositional nature (Davies et al., 2011). For example, it is entirely possible to imagine that everyone is excited without creating a mental image of it. Kosslyn (1994) refers to these proposition-like entities, which describe a scene, as “mental sentences” or descriptive representations. Goldman (2006) suggested different types of imagination, such as S-imagination (supposition imagination) which is usually formulated with a “that”-clause and characterized by being completely conceptual and therefore lacking any sensory component. For example, one could imagine that “ $x = 2$ ” and “ $y = 5$ ” or even that “the dog is on the bed”. E-imagination (enactment-imagination) was also suggested by Goldman (2006) and involves sensory forms of imagination (not necessarily reliant on imagery) to recreate perception-like states. For example, one could imagine they were enacting the motions necessary to kick a soccer ball without the intention of actually executing this action.

Visual imagery is measured with subjective reports of how vivid an individual’s imagery is (Keogh & Pearson, 2018). For example, when asked to imagine the face of a close relative, some people claimed imagery so vivid that it was almost as if they were seeing that person and others recounted seeing no visual image at all, despite knowing that they are thinking about that person (Keogh & Pearson, 2018). Many others fall somewhere in between these two, and as such these results are better represented by a spectrum of vividness, rather than binary (image or no image). The inability to generate mental images is known as aphantasia (Zeman, Dewar, & Della Sala, 2015).

This phenomenon was initially described by Galton almost 150 years ago, but has only gained public and scientific interest

recently (Jacobs, Schwarzkopf, & Silvanto, 2018). This inability to experience mental imagery presents us with a new model to study the functional role of mental imagery in visual cognition. Research on the functional role of mental imagery in visual working memory from Jacobs et al. (2018) indicates that it is possible to compensate for a lack of mental imagery under some conditions by using different cognitive strategies to achieve similar results. It is possible that subjects use propositional semantic information about images rather than creating a visual image in their mind (Keogh & Pearson, 2018). Instead of using low-level visual object imagery it is possible to perform mental rotation tasks using spatial or kinesthetic imagery (Keogh & Pearson, 2018).

Aphantasics are unable to experience the sensory qualities of objects that are beyond their perception (Jacobs et al., 2018). This does not mean that they are incapable of imagination or dreaming. Compared to a control group of participants with visual imagery abilities, aphantasics reported less vividness in autobiographical memories and in imagined future scenarios as well as having fewer (and less vivid) dreams (Dawes et al., 2020). Despite the loss of quality and vividness in other aspects, spatial abilities remained unaffected by a lack of mental visual imagery (Dawes et al., 2020).

In spite of the wealth of studies about the imagination in general, there is very little work on how objects are placed in imaginary scenes. This work attempts to describe, and implement in software, a model of how this might be done in human visual imagination.

While our mental representations are based in part from our perceptions of reality, imagination is not necessarily realistic (e.g., parts of scenes can be incomplete, or fade away) and we are likely to have more stereotypical representations in our mental representations than those present in reality. Although my intention is to make this model imagine scenes in a psychologically plausible manner, I must first address the most noticeable issues, which happen to be the same issues I would need to address if my intentions were to model reality (e.g., gravity, object scaling, object positioning).

2 Previous model

This section discusses a previous model of imagination and motivates the implementation of a new approach to the scaling and placement of objects in a 3D scene to better simulate human imagination. I describe the Science of Imagination Laboratory’s Imagination Engine (SOILIE) and the steps involved in creating images of 3D scenes. I also showcase sample output and highlight some of the more notable issues that differentiate these images from scenes that could plausibly be imagined by humans.

2.1 Science of Imagination Laboratory’s Imagination Engine: SOILIE

The focus of the Science of Imagination Lab (SOIL) is on understanding and explaining human visual imagination with the aid of artificial intelligence. More specifically, the lab’s collective objective is to create computer programs that can take in minimal input and create detailed visual scenes. SOILIE (Science of Imagination Laboratory’s Imagination Engine) is a Python computer program designed to imagine scenes in 2D the same way that humans do (Breault, Ouellet, Somers, & Davies, 2013). Taking in only a single word inputted by the user, SOILIE can generate a 2D image of a scene including the query word as well as several associated images, in mostly correct places. Although SOILIE can generate images with coherent content and placement in 2D, these images are rather far removed from our internal experiences of imagery (Breault et al., 2013). Further work is being done to improve the ability of the engine in realistically modelling human visual imagination.

SOILIE 3D was designed by Vertolli and Cichonski for SOIL with the intention of modelling 3D scenes similarly to how humans imagine such scenes. While SOILIE 3D is an intellectual descendant of SOILIE 2D, it is a separate program and shares no input nor output. Comprised of Python scripts written by Cichonski, SOILIE 3D can be broken down into three main steps: (1) collection of data, (2) preparation of data, (3) placement of objects and rendering a final 3D scene.

2.2 SUN3D database as input for SOILIE 3D

SOILIE 3D requires the data gathered from the SUN3D database as input. The SUN3D database is a large-scale RGB-D (Red Green Blue – Depth) video database and, unlike most existing view-based databases, it is intended to be a place-centric and a complete representation of 3D spaces (Xiao, Owens, & Torralba, 2013).

Data was captured by operators who walked through rooms with an ASUS XTION PRO LIVE sensor taped to a laptop, which they held at head level (and with a slight downward tilt) (Xiao et al., 2013). The goal was to imitate how a human would explore the scenes; the operators panned the rooms moving slowly up and down, and side by side. Only indoor spaces (classrooms, labs, offices, apartments, etc.) were scanned due to the limitations of the depth cameras (they do not function under direct sunlight; Xiao et al., 2013).

The SUN3D database is structured in a way that each frame for each scene is stored separately. The JSON files are lists containing the name of the scene scanned, the date the scan was performed, the frames, and the objects and polygons represented. Figure 1 shows part of a CSV file for one scene within the database and Figure 2 is an image of one of the frames from one of the scenes within the SUN3D database. Objects were manually labeled by humans.

```
[{"date": "Tue, 03 Dec 2013 00:07:03 GMT", "name": "hotel_umd/maryland_hotel3", "frames": [{"x": [166.5, 89.5, 96.5, 102.5, 119.5, 149.5, 178.5, 175.5], "y": [352.5, 341.5, 313.97, 2.5635221004486084], [-1.108713150024414, 0.18478552997112274, 2.5093092918305996], [-0.915153980255127, 0.1821834295988083, 2.4164395332336426]], "object": 1}, {"x": [516.5, 407.491789460182, -0.055814679712057114], [-0.015164893120527267, -0.005007941275835037, -0.04022873938083649], [-0.016257109120488167, -0.005935135297477245, -0.036794111132621765], 1936035], [-1.2199347019195557, 0.15755614638328552, 2.5674548149108887], [-1.3114317655563354, 0.24302756786346436, 2.6152617931365967], [-1.3200223445892334, 0.0365404114127159193, -0.720327615737915], [-0.06252036988735199, -0.2637578248977661, -1.1143126487731934], [0.008118228986859322, 0.01047513447701931, 0.1493602842092514], [0.005118473432958126, 0.97845], [0.05107550323009491, 0.017860759049654007, 0.1787148267030716], [0.030805159360170364, 0.012652118690311909, 0.15687038004398346], [0.023757901042699814, 0.0156381111592626465], [-0.38081324100494385, 0.4714111387729645, 1.7223937511444092], [-0.3677423298358917, 0.6637300848960876, 1.7051949501037598], [-0.44418469071388245, 0.7271242737770081, 1.016851479187607765], [-0.0007836183649487793, -0.004278132691979408, -0.012079206295311451], [-0.0009049161453731358, -0.0032975757494568825, -0.00874765869230032], [0.0356477
```

Figure 1: Sample CSV from one scene in the SUN3D database

Figure 2: (2013).[Frame from one scene in the SUN3D database]. SUN3D Database. <http://sun3d.cs.princeton.edu/>

2.3 Step 1: collection of data

SOILIE 3D uses the original, manually labelled data from the SUN3D videos as input. The first step is to read data from the SUN3D database (JSON files), find the centroids of the polygons, and calculate the angles and distances between object triplets (A, B, and C) in each frame. A CSV file containing the calculated distances in metres (AB , AC , AO) and angles (OAC , OAB , BAC) between their respective centroids is then created. Note that O represents the observer (or camera). Figure 3 is a sample CSV created for one of the frames within one of the scenes in the SUN3D database.

```

objectA,objectB,objectC,distanceAB,distanceAC,distanceAO,angleOAB,angleOAC,angleBAC
cabinet_1,chair_1,table_top_1,1.26212002445,1.63233528241,3.22470150773,15.5146668412,20.7077936399,31.7737918799
cabinet_1,chair_1,chair_2,1.26212002445,0.636436350175,3.22470150773,50.3641319868,20.7077936399,63.3180900403
cabinet_1,chair_1,chair_3,1.26212002445,1.36135494341,3.22470150773,136.780946848,20.7077936399,143.814316013
cabinet_1,chair_1,chair_4,1.26212002445,1.8794414286,3.22470150773,92.637001825,20.7077936399,104.542105296
cabinet_1,chair_1,chair_5,1.26212002445,2.70464685366,3.22470150773,81.5873313704,20.7077936399,95.3230174019
cabinet_1,chair_1,opaque_window_1,1.26212002445,2.44929077852,3.22470150773,132.372432342,20.7077936399,131.582275082
cabinet_1,chair_1,wall_1,1.26212002445,0.893338433475,3.22470150773,143.677032732,20.7077936399,142.330984271
cabinet_1,chair_1,wall_2,1.26212002445,0.947952756987,3.22470150773,110.198264735,20.7077936399,96.567874318
cabinet_1,chair_1,box_1,1.26212002445,0.264394532656,3.22470150773,91.6021462173,20.7077936399,76.5680476494
cabinet_1,chair_1,closet_1,1.26212002445,2.52838557026,3.22470150773,124.116736744,20.7077936399,135.011022411
cabinet_1,chair_1,floor_None,1.26212002445,0.495220134715,3.22470150773,106.169167638,20.7077936399,126.024432313

```

Figure 3: Sample CSV created from one frame within one scene in the SUN3D database

2.4 Step 2: preparation of data

Step 1 collects all the distances and angles from all the triplets but it is separated by frame and scene. In order to prepare the data, Step 2 takes the CSV files created from Step 1 and it averages the values of the object triplets, and estimates the sizes of objects.

Unfortunately, the 3D data is incomplete, which impedes the calculation of the true volume of objects. Relative object sizes are therefore estimated based on the object's largest measured diameter. The resulting output is a CSV file of triplets with calculated angles and distances between the permutations (see Figure 4) as well as one containing calculated object sizes in metres (see Figure 5). As shown in Figure 4, the output from Step 2 contains scalar distances between object triplets (e.g., distance from the wall to the chair is 2.113516).

```

objectA,objectB,objectC,distanceAB,distanceAC,distanceAO,angleBAC,angleOAB,angleOAC,count
chair,chair,chair,2.0854309621715843,2.0854309621715843,4.132925289609801,60.00000000000033,80.42401922445576,80.42401922445566,91374
wall,chair,wall,2.050664110735824,1.8036759184590954,4.830501890993017,62.45063984255222,62.664396626648234,82.02959692691711,3520
wall,chair>window,2.243107089076964,1.8027337306253628,5.427880256920235,80.75208407737325,57.25506462577917,96.50162935123961,1580
wall,chair,chair,2.1135165511630727,2.1135165511630727,4.978564560916499,52.797359553838476,60.3322746436236,60.332274643623606,17014
wall,chair,white_board,2.1351971646176295,2.074785675611042,4.981507731288977,66.22145865444851,64.44025263896239,69.70569255410335,615
wall,chair,floor,2.0144104300814285,2.328210364865448,4.743121388772741,48.815249967106475,62.79117110960361,53.86393019571144,1722
wall,chair,table_top,2.135047907131121,2.4618485613080616,5.028486775465488,48.1122025416655,63.17537988082347,39.59837585094455,1958
wall,wall,chair,1.8036759184590954,2.050664110735824,4.830501890993017,62.45063984255222,62.0295969269171,62.664396626648234,3520
wall,wall>window,1.8461789120917642,1.8221744243170694,4.479375544841496,61.64783748014362,81.57288963367006,99.98073819941888,632
wall,wall,chair,1.8036759184590954,2.050664110735824,4.830501890993017,62.45063984255222,62.0295969269171,62.664396626648234,3520
wall,wall,white_board,2.0827870896336322,2.284689301976425,4.919572463761333,66.83721348955275,80.58450198207566,57.7987737431376,322
wall,wall,floor,1.5817821768707332,1.7634092441419496,3.9039656934442886,63.70773608220822,81.1110019580713,70.64572578933387,958
wall,wall,wall,1.6508424695151787,1.6508424695151789,3.7812080861016346,60.00000000000013,80.92647602435389,80.9264760243539,2316
wall,wall,table_top,1.088672717607154,2.44013430344348873,4.874040438277088,67.0450032157083,82.06660670766651,43.44111871728405,538

```

Figure 4: Sample output of triplets

```

object,diameter,count
wall,19.44036713628508,876
floor,15.80952618098821,268
black_board,2.6363774576235537,9
table_top,12.486367178085223,200
door,19.868833701428688,110
cabinet,15.883421375189712,54
closet,4.815413055117433,26
chair,6.996521900389839,1264
opaque_window,12.197996637659113,51
box,2.124411352524028,6
window,12.226112028572933,101
clock,1.484425614587969,3
white_board,15.96441360887978,111
wooden_wall,12.04820281242249,14
window_sill,3.202384993721975,1

```

Figure 5: Sample output of size

2.5 Step 3: placement of objects and rendering a final 3D scene

The final step is the imagination phase. This is run using Blender in the terminal and requires the output from Step 2 as well as the 3DS files with matching object names obtained from the SUN3D database.

This step essentially takes a list of objects (randomly chosen or selected by the user) and determines their Cartesian coordinates based on the angles and distances between triplets. It generates and goes through every permutation of said objects in order to find a set of triplets containing enough data to calculate all the object locations. Based on the objects inputted, a list of the triplets required for the calculation of final object positions is created. If any of these objects are not in the final list, this means that there is not enough data, and the program will exit. Otherwise, the program will continue.

Triplet data is found for each combination, and from Step 2 the scalar distances between A and B , A and C , and A and O , as well as the angles BAC , OAB , OAC are used to determine the position of the objects in a Cartesian coordinate system.

Table 1 shows the known variables taken from Step 2 and the values that they represent.

Variable	Description
d_{ab}	distance between object A and object B
d_{ac}	distance between object A and object C
d_{ao}	distance between object A and camera
$aBAC$	angle BAC
$aOAB$	angle OAB
$aOAC$	angle OAC

Table 1: Distances and angles

Using the imported distances, angles, and centroid locations for A , B , C , and O , the objects are placed within the scene. Afterwards, a scale of objects is created so that they can fit into the camera's range of sight.

Table 2 shows the centroid locations for objects A , B , C , and O .

Object	Position
A	$(0, 0, 0)$
B	$(d_{ab}, 0, 0)$
C	$(a, b, 0)$
O (Camera)	(x, y, z)

Table 2: Centroid locations

Object A is arbitrarily assigned the origin coordinate $(0,0,0)$. Now object B must be placed. Object B is placed at distance d_{ab} from object A and along the $xaxis$ so that object B is at $(d_{ab},0,0)$. In order to place object C , we must calculate the values for a and for b . Object C is a point in the $xyplane$ and we can therefore calculate the value of a and b using simple trigonometry. Distance d_{ac} and angle BAC were already known from Step 2 and therefore we can use the following equations to find the x and y values of C position.

$$a = d_{ac} \cdot \cos(BAC) \quad (1)$$

$$b = d_{ac} \cdot \sin(BAC) \quad (2)$$

Object O (camera) is in another plane, with coordinates $O(x,y,z)$. Since d_{ao} and angle OAB are already known from Step 2, we can use the following equation.

$$x = d_{ao} \cdot \cos(OAB) \quad (3)$$

The radius of a sphere is $r^2 = x^2 + y^2 + z^2$ and if we take d_{ao} to be r , we can isolate z and use the following equation to calculate the z value of object O .

$$z = \sqrt{(d_{ao})^2 - x^2 - y^2} \quad (4)$$

Knowing that the dot product of two vectors is equal to the length of the first one multiplied by the length of the second one and the angle between them, we can isolate for y and use the following equation to calculate the y value of object O .

$$y = \frac{(d_{ac} \cdot d_{ao} \cdot \cos(OAC) - a \cdot x)}{b} \quad (5)$$

An arbitrary scale factor is implemented to ensure that all of the objects will be within the observer's range of vision; objects were scaled down so that they were all visible by the camera.

2.6 Output

When running SOILIE 3D from the terminal, the user is presented with a menu (see Figure 6).

```

SELECT AN OPTION
-----
1. Imagine a random scene
2. Enter object names
3. Determine working combos
4. Exit

```

Figure 6: SOILIE runtime menu

If choosing menu option 2 ("enter object names"), the user can select specific objects to be within the scene. If the user wants to input a custom list of objects, then they need to enter object names that already exist in the database and that have sufficient triplet data. There is a file (see Figure 7) containing a list of all currently working combinations.

```

back_pack,bed_frame,blanket,ceiling,wall,
back_pack,bed_frame,blanket,ceiling,,
back_pack,bed_frame,blanket,chair,ceiling,wall
back_pack,bed_frame,blanket,chair,ceiling,
back_pack,bed_frame,blanket,chair,helmet,ceiling
back_pack,bed_frame,blanket,chair,helmet,wall
back_pack,bed_frame,blanket,chair,helmet,
back_pack,bed_frame,blanket,chair,power_socket,ceiling
back_pack,bed_frame,blanket,chair,power_socket,helmet
back_pack,bed_frame,blanket,chair,power_socket,wall
back_pack,bed_frame,blanket,chair,power_socket>window_shade
back_pack,bed_frame,blanket,chair,power_socket,
back_pack,bed_frame,blanket,chair,wall,
back_pack,bed_frame,blanket,chair>window_shade,ceiling
back_pack,bed_frame,blanket,chair>window_shade,helmet
back_pack,bed_frame,blanket,chair>window_shade,wall
back_pack,bed_frame,blanket,chair>window_shade,

```

Figure 7: Sample of working combinations

If choosing menu option 1 ("imagine a random scene"), the user can choose to have a random scene generated by specifying the number of objects to be contained in the scene. Note that this script requires a minimum of three objects to function. In theory there is no limit to the number of objects that can be placed in a scene using this method. In practice, due to a lack of data, SOILIE 3D can place a maximum of six objects (including the camera) in a scene. All output I have presented in this thesis has been created using this random method.

The output generated by SOILIE 3D is a PNG image (see Figure 8).

Figure 8: Sample image output of a generated scene

From this, a corresponding CSV file (see Figure 9 below) containing object labels and the position (x,y,z coordinates) of each object centroid can be generated. The CSV file also includes a field containing the size of each object as well as the scale factor used to resize all the objects (actual object size = size * scale). Note that objects are resized to allow them to all be visible in the camera's field of view and that size units used thereafter are arbitrary. Actual object sizes are estimates of diameter (in metres) calculated from SUN3D data.

```
object,x,y,z,size,scale
CAMERA,-1.4126998265789683,-1.0692825925567893,4.022856166750316,0,0
dining_table,0,0,0,0.33462527200995384,1.582145581053907
lamp,1.4125486749089466,-0.2628339876485536,-1.2003360753440062,0.23407752385075026,1.582145581053907
chair,0.6698213189101035,1.606091270832095,-2.249973027779234e-08,0.3500900685044041,1.582145581053907
floor,-1.7882677379422396,26.938532774858558,-54.675515325727545,0.3955363929271698,1.582145581053907
table,-0.6344179263403292,-0.7327958681737154,1.3451724930969533,0.18514060271285454,1.582145581053907
plant,2.4179056875565665,0,0,0.34046443439824864,1.582145581053907
```

Figure 9: Sample CSV output of a generated scene

2.7 Notable Issues

While SOILIE 3D is fully capable of generating novel 3D scenes, there is something not quite right with these scenes; the scenes generated by SOILIE 3D do not fit with human experiences of imagined scenes. While technically there is nothing wrong with imagining all the furniture to be clustered together or floating haphazardly in space, it is extremely doubtful that a human would do so, and it is not an accurate representation of how a human would imagine a scene with the same objects. The same can be said for size; trash cans double the size of a sofa or a table seem like an unlikely choice.

For example, if I were to ask someone to imagine a scene containing a table, chairs, a mug and a painting, although the scene would vary from each person, there are some key aspects which would remain the same in most instances. The issues with the images currently being generated by SOILIE 3D is that there are no methods to scale objects appropriately (mugs should be smaller than chairs), or understand that certain things should be facing specific directions (chairs should face the table), or even understand that most objects are supported in some way (tables touch the ground, paintings are usually on walls, chandeliers hang from the ceiling, etc.). People do not need to be told these things. It is part of our understanding of the world and how objects function within it; it is just common sense. People do not consciously think of these regularities when placing objects into the scene, it just naturally happens.

2.7.1 Data limitations: Small total number of objects

The scenes and objects in SOILIE 3D are dependent on data collected from SUN3D. The distribution of places and categories that were scanned for SUN3D were disproportionately skewed towards office and school settings and due to camera limitations did not include any outdoor scenes. The data in these scenes is also incomplete due to missing labels, as well as requiring sufficient data to calculate necessary triplets for SOILIE 3D. All of this contributes to limitations in the choices of the objects SOILIE 3D can place in a scene. While this is mostly due to the limitations of the SUN3D database, the availability of variations of actual 3D models of the selected objects leaves much to be desired - we can only put objects into 3D scenes if we have models of those objects. While the human imagination is vast, and we can easily come up with countless variations and combinations, SOILIE 3D is hampered by these restrictions.

SOILIE 3D selects objects to go in a scene together based on the triplet data; if there is enough data, that object is selected, if not, it moves on to the next object. The problem with this approach is that even if these objects are together in a scene and have enough data, sometimes they don't make sense as a subset alone. For example, some scenes include windowsills without windows or a window without a wall to anchor it. Other scenes might include a desk lamp, without a plausible piece of furniture to rest on or multiple tables without any chairs.

2.7.2 Object parts versus wholes

Another issue is that, for some objects, SOILIE 3D takes labels of object parts and does not treat them as a whole, whereas almost all other objects are treated as a whole. For example, a desk is treated as one object whereas a table is not always treated as one complete object (table, table_top, table_leg) This is also problematic because some components of these objects are listed (bed_frame and bed) while others are not (mattress). This causes a scene to be made up of five "objects", while in reality it is mostly made up of the components for one object (a table, a table top, a table leg, etc.).

2.7.3 Object placement

For the purposes of this thesis, I assume that the objects have been selected appropriately and focus on the next step in the imagination process. Other than the issues mentioned above, there are notable issues with the size, placement, and rotation of objects. While Figure 8 is one of the more successful images created by SOILIE 3D, Figure 10 shows more typical output from SOILIE 3D and highlights the current problem. Most of the images had obvious issues with both scaling and placement of objects.

Figure 10: Problematic output from SOILIE 3D

One of the most notable issues is object placement. SOILIE 3D does not account for gravity and objects are almost always positioned at extreme distances either above or below the floor. The distance between some objects is also problematic, as some objects are placed so far out of the scene they seem to belong to a whole other scene, or are positioned above other objects and overlapping. Since it was believed that the estimates of object sizes and distances between the objects were relatively accurate, there is nothing in the original code from SOILIE 3D that would prevent objects from overlapping.

From our daily experiences interacting with objects in the world, we have gained some knowledge of physics; we know that gravity exists and solid objects do not pass through other solid objects. We expect objects are supported on a horizontal plane (e.g., floor) or on a vertical plane (e.g., wall). A ceiling and objects such as pendants or chandeliers should be on another horizontal plane above the floor. Other objects should fit somewhere within the bounds of the floor and ceiling. At no point should these objects appear below the floor or intersecting with each other.

2.7.4 Object scaling

As can be seen from the output images (Figure 8 and Figure 10), objects are scaled inaccurately. In Figure 10 (iii), the trash bin is the biggest object in the scene and even the vase is larger than the table. In Figure 10 (iv), the trash bin and desk lamp are bigger than the dining table, chair, and sofa.

The problematic scaling is due to the size of object models and the arbitrary scale factor used by SOILIE 3D to ensure visualization by the camera.

3 Improved model

This section discusses the methods I used to improve the previous version of SOILIE 3D and details the research supporting these modifications. I begin by describing the refactoring and updates needed to run the previous (and outdated) model. I then explain the steps necessary for the creation of the before images as well as those necessary for the after images. Although there is not a lot of research on how humans imagine scenes, by extending the findings of research on human scene understanding towards scene imagination, I developed a more plausible approach to the scaling and positioning of objects within a 3D scene. This software demonstrates the multiple constraints that must be satisfied to simulate a plausible model of imagination. It is, however, not a process model, as the exact manner in which these constraints are implemented is as of yet unknown. The creation and comparison of 30 pairs of before and after images can further help inform the important features that humans consider when imagining scenes.

3.1 Improving SOILIE 3D output

Whether our interactions in space are real or imagined, the mental representations we form of space are different from the external representations of space in terms of physics, geometry, or maps (Tversky, 2003). The previous model was too theoretical and did not take human cognition into account, causing noticeable discrepancies with the placement and scaling of objects within the scene.

The first step in the visual imagination process is creating the descriptive representation of it, which is aptly named “scene description” (Davies et al., 2011). In order to imagine a scene, one requires access to both episodic (events) and semantic (facts) memories. Despite containing a lot of visual and spatial information, these memories are stored as descriptive representations; they are structured descriptions, or sentence-like structures consisting of symbols. While the precise code and exact symbols used in our mind’s descriptive representations are unknown, scene descriptions are typically structured as follows (Davies et al., 2011).

House

Brick(house)

In_front_of(garden, house)

Has(garden, flowers)

While SOILIE 3D is a good first step in the imagination process, it has not been updated since 2018. It was written in Python 2 and Python 3 and made use of certain Blender modules and 3DS files that are now outdated. I used output from a slightly modified SOILIE 3D version to determine which objects will be placed within the scene and their initial positions. I refactored the original SOILIE 3D code all to Python 3 but was unable to successfully run the program through Blender due to compatibility issues. Therefore I modified Step 3 (placement of objects and rendering) to directly generate a CSV file instead of a PNG and implemented the imaging aspect through Unity with different 3D models. The 3D object models were taken from the free Unity Asset Store items. They were not all from the same designer and therefore have different scales. The previous version of SOILIE 3D had an arbitrary scale set to objects in order to properly fit them into the scene; objects would be scaled by a different factor in each scene in order to be viewed within the camera. Since this was an arbitrary number and different models are being used, scale information from SOILIE 3D output was not used for the new images. Due to the limited availability of free and properly rendered 3D models, I was unable to find models for all of the objects in the SOILIE 3D item list. I created 2D Sprites to fill in until such models can be found or created.

I used the output from my slightly modified version of SOILIE 3D as the scene descriptions required to imagine a scene (see Figure 11 below).

```

window,"[0, 0, 0]"
wall,"[1.1977632162157856, 0, 0]"
stand,"[1.280576160216201, 1.081740115325029, 0]"
CAMERA,"[-0.5732510281155807, -1.5335566243882897, 1.8943159293828278]"
light,"[-0.2384080678741313, -1.2485909244654654, 3.032140205745451]"
door,"[1.29141932039214, -9.404767769174521e-09, 2.46871235214462]"

```

Figure 11: Sample CSV files used for creation of following scenes

The 30 pairs of before and after images of each scene can be found in Appendix A.

3.2 Before scenes

I will describe what the original SOILIE 3D produced in terms of “before” scenes, and those that are created after my modifications “after” scenes. Scene descriptions were always chosen as 6 random objects (1 being the camera) so that a total

of 5 objects would be present in all of the scenes. When generating scene descriptions, those involving many components of objects (table_top, table_leg, table, etc.) rather than complete objects were excluded.

Objects were positioned using the exact cartesian coordinates determined by SOILIE 3D scene descriptions. It should be noted that that previous SOILIE 3D objects were positioned by centroid locations and the objects in the current version are positioned by pivot point. This is due to the way Unity 3D models are typically created; the current models in Unity are instantiated by pivot points which are by default at the bottom right corner of object models.

Figure 12: Before image of Scene 1

3.3 After scenes

While imagined scenes are bound by the knowledge constraints of the individual imagining it, there are certain common physical properties that must be respected. To create the after scenes, objects were instantiated using the same scene descriptions as the before scene. Changes were then made so that objects were no longer intersecting with others, objects were no longer floating in space, and objects were all sized in plausible ways in respect to each other. The following subsections detail the changes made and the research supporting these decisions.

Figure 13: After image of Scene 1

3.3.1 Object placement

Humans can make incredibly robust and quick inferences, even in the absence of crucial information, and this is in part due to our imagination. Visual imagery has many important uses, especially in reasoning and problem solving when the stimulus is no longer present in the visual environment.

Object placement might not be as random and unstructured as one might think. Even if one thinks that they are placing objects into a scene haphazardly, there are still constraints by which we abide. Despite everyone imagining something slightly different, there are some commonalities that make it a plausible scene (e.g. chairs will probably face the table, and they will all be on the floor). Without much thought, we know what is free to be lifted, what is firmly attached, what might be knocked over, and what should be in motion. Although it takes but a moment to know all this from an imagined or perceived scene, we have yet to understand all the mechanisms at work behind this intuitive physics.

Battaglia, Hamrick, and Tenenbaum (2013) proposed a simulation-based intuitive physics engine (IPE) model to explain people's physical scene understanding. Their IPE model explains core features of human mental models and common-sense reasoning which are essential to how humans understand the world around them and it makes robust and fast inferences by using approximate, probabilistic simulations in complex natural scenes where crucial information is absent. The IPE model offers a plausible algorithmic basis for how humans can quickly understand scenes around them; probabilistic simulations can help explain how perceived scene layouts, causal inferences, movement planning, language semantics and syntax can all be

influenced by our physical knowledge (Battaglia et al., 2013).

While an intuitive physics model would have been ideal, I believe that it would be too complex for the scope of this project. As such, I focused on implementing a simplified understanding of physics, and human judgments by constraining degrees of freedom and object placement. In order to realistically simulate a scene that a human could imagine, I attempted to determine the constraints that humans are bound by. There is not much research on how humans imagine scenes. I therefore focused my research on how humans understand scenes and made the assumption that properties of importance in understanding scenes would be of equal importance in generating them.

We need mental representations of space in order to act effectively in them, and these are quite different from a physicist's or cartographer's representations of space (Tversky, 2003). Our interactions in a space can be explicit (grabbing a nearby mug), implicit (being aware of our surroundings), or imagined (estimating the distance from your house to the nearest St-Hubert). Our mental representations are created from knowledge obtained from various sources (sight, touch, sound, imagination, language). All of these knowledge sources can support spatial activities, and for many actions, the source of knowledge is interchangeable. For example, we can get to St-Hubert while using a GPS or verbal directions from our mother.

Mental representations are based on the elements in the space and the spatial relations between them relative to a particular reference frame. Depending on the task at hand, people act in different spaces, such as the space of the body, the space surrounding the body, the space of navigation, and the space of graphics. Our mental representations will then differ based on which elements and spatial relations are "central for functioning in the different spaces" (Tversky, 2003).

In geometry, physics, and cartography, space is typically measured based on a uniform and unitary metric value (Tversky, 2003). This is part of the problem with the initial SOILIE 3D output. It is using metric and geometric representations of space instead of using a more relativistic approach. In human conceptions of space, spatial relations amongst things in the space, with respect to a reference frame, form the framework for mental spaces. The items to be selected and the determination of relevant spatial relations depend on the space they will fill as well as the functions it serves us. All spaces are represented schematically in terms of the objects located within and the spatial relations that are necessary for functioning within it (Tversky, 2003).

In order to position an object in a 3D scene, one should technically control for six independent variables (3 for positioning, and 3 for orientation) (Smith, Salzman, & Stuerzlinger, 2001). Although this is how SOILIE 3D initially placed objects, I believe it to be problematic. Some objects have much less variability in their possible placements than others, which should be considered. For example, a picture frame has many placement possibilities (on walls, leaning against other objects, placed on other surfaces such as desks, tables, etc.) whereas a refrigerator only has a limited number of plausible placement options (must be against a wall and not in the center of a room). Observations of humans moving furniture and designing environments have shown that humans do not think about scene manipulation as a problem with six degrees of freedom (Smith et al., 2001).

This can be explained by the constraints that human conventions and physics place on objects (Smith et al., 2001). We expect objects to be right-side up and obey the laws of physics; beds do not float, lamps are usually plugged in and upright, and picture frames are anchored to a wall or leaning against a surface. This can explain why a lot of the initial SOILIE 3D output was technically correct, but one would feel a large disconnect while looking at the scenes. While *theoretically* most objects have six degrees of freedom, *in practice* most real objects have a maximum of three degrees of freedom and rest on a plane (Smith et al., 2001). Many of these objects are actually placed against walls or other objects, which reduces the available degrees of freedom even more (Smith et al., 2001). This was one of the main problems with the previous SOILIE 3D version: it focused too much on theoretical calculations and missed the practical reality of how humans view and think of space and objects that fill it. For example, in one of the scenes, SOILIE 3D positions the kitchen counters not on a wall and a bookcase between the counters and the wall. No one would place a bookcase in front of the counters because it disrupts the functionality of the space. One needs to be able to move about in a kitchen and open cupboards and a bookcase should be against a wall, not blocking you from getting a bowl or cooking supper.

The manner in which an object interacts with the physical world can assist a user in placing and manipulating objects in virtual environments (Smith et al., 2001). To simulate this, objects in a scene are given a set of rules (constraints) which must be followed when they are being manipulated (Smith et al., 2001). Since it is inefficient and an incredible undertaking to bestow human-like common-sense upon a computer, I have implemented what I believe to be some of the more important constraints to objects being placed in a scene. More specifically, I have implemented constraints for the surface against which objects should be placed against (vertical vs horizontal) and limitations on the x, y, and z axes. Objects were set to rest on the ground to respect gravity. Objects that were not restricted to the ground were those that were typically against a vertical surface (windows, walls, doors, clocks, plaques, etc.) Some of these objects were set below the surface of the ground (for all scenes ground is set to $y = 0$). For any of these objects set below the surface, the value of y was set to 0 (remember that objects are instantiated by pivot point which is at the bottom right corner of the 3D model). Some objects were set too high above the others, so a maximal y value of 5 was set. Otherwise, the y value remained unchanged.

When positioning objects in a 3D scene, Bukowski and Sequin (1995) used a combination of pseudo-physical and goal-oriented properties called “Object Associations”. Gosele and Stuerzlinger (1999) built upon these ideas and gave each object in the scene predefined offer and binding areas, which were used to define the constraining surfaces between objects. A mug, for example, has a binding area at its base, whereas a coffee table has an offer area on its top. Therefore, a mug can be constrained to a tabletop. To better simulate how objects behave in the real-world, a labeled hierarchy adds semantics to the constraint process (Smith et al., 2001). Associating each constraint area with a label from this hierarchy restricts a binding area only to offer areas with labels equal to or descendant from in the hierarchy. This allows the legs of a chair to be constrained to the floor, or a desk, but never to the wall (Smith et al., 2001). I have implemented some rules to determine whether objects

can be placed upon another or not. Objects were manually categorized by binding areas. Figure 14 shows a sample of objects and their possible binding areas (top, bottom, both, no).

door	no
cabinet	top
closet	no
chair	top
box	both
window	both
clock	bottom
white_board	bottom
window_sill	both
dust_bin	bottom
bin	bottom

Figure 14: Object binding areas

Objects were then inspected to see if they were intersecting with each other. Since this was a problem in every scene, objects were spaced apart by an arbitrary scale factor of 2.

Afterwards, if objects were intersecting, or under others, they were tested to see what their attachment points were. I manually added "bottom", "top", or "both" to objects as attachment points. If an object with "bottom" was below another object that was either "top" or "both", then it was placed on top of the former. Finally, if an object was still intersecting with another, it would be shifted by small increments along the horizontal plane on which it lies until it was no longer intersecting. Objects such as walls, windows, blinds, etc. were allowed to intersect (windows can be set on walls).

Figure 15 shows an image where a chair is allowed to be below a table.

Figure 15: Chair below table

The objects are not intersecting and based on the binding constraints, the chair can remain in its current position. If objects intersected, or were vertically aligned, and there were no acceptable matches of binding areas for these objects, then they would be shifted along the x and z axes until they were no longer intersecting. Figure 16 is a flowchart showcasing the decision process for dealing with intersecting objects.

Figure 16: Decision flowchart for intersecting objects

There were also issues in some of the 'before' scenes where one object would be set at an incredibly far distance from the rest of the objects. This had a jarring effect on the 'wholeness' of the scene since it gave the impression that the objects were not of the same scene and did not belong together. Figure 17 is an example of a scene with objects placed unnaturally far apart.

Figure 17: Problematic before scene with objects placed unnaturally far apart

To fix this issue, objects were restricted to have a distance of no more than 50 units from other objects. Based on the standard distances between objects within all the scenes, 50 units was chosen as a reasonable maximal distance. Figure 18 shows the improved image of this scene. While this is still not an ideal representation of a scene, it can easily be considered a scene while Figure 17 can not.

Figure 18: Improved scene with objects placed closer together

Another issue with previous object placement was that objects that should be restricted to the vertical boundaries of a scene (such as windows, doors, and blinds) were placed in the middle of a scene. Figure 19 shows an example of this issue.

Figure 19: Problematic scene with blinds placed in center of a room

In order to fix this issue, objects that should be placed along walls were restricted to the boundaries of the scene. While this did improve the issue, these objects did not always appear in ideal locations. Figure 20 shows the improved version of the previous figure. While the blinds are now correctly placed along the bounding box, it does look peculiar without a wall.

Figure 20: Improved scene with blinds placed along the bounding box of the room

Other scenes had walls and objects that should be placed on the walls. While the improved images successfully had them on the same plane, their exact locations were not always ideal. Figure 21 shows an example of objects that were correctly placed along the bounding box of the room. The scene is still problematic since the wall does not fully extend to where the edges of the bed are and the painting is not fully on the wall. While these objects were placed according to scene descriptions from SOILIE 3D and respecting the new constraints I implemented, there is still room for improvement. For instance, if given blinds and a window as objects within a scene, it is more likely that humans would place the blinds on top of the window as that is acceptable placement of blinds that takes into account its functionality. While there is technically nothing wrong with placing blinds against a wall, it just doesn't make much sense to do so.

Figure 21: Improved but still problematic window and blinds placement

3.3.2 Object scaling

Humans do not use exact magnitudes when creating mental representations of objects; these representations are more vague and should account for perceptual uncertainty. These magnitudes should be represented as distributions of fuzzy sets (Gagné & Davies, 2013). We do not think of tables as being 28-30" high, just as we do not think of cups as being 6" tall. We also do not think of objects as being stored in broad size categories, however, due to the limitations of this thesis, this is the method applied for scaling objects in the current model. Unfortunately, there is no universal 3D object database with a constant scale and in order to properly represent these objects in a plausible scene, a common scaling system must be developed and applied. Objects were resized using a XS, S, M, L, XL category grouping. I manually inspected items and took average measurement (in inches) of all objects. Objects were then manually assigned to categories and resized accordingly. Unlike the previous version of SOILIE 3D, scaling of objects was consistent throughout all scenes. This is not the most accurate representation of human understanding of size. This part of the model is not intended to model human psychology; it is a rudimentary stopgap.

There are regions of the visual cortex that respond differently to large and small objects and this difference is most likely because of navigational processes; we can move small objects, but in order to navigate around larger objects that are too big to be moved, we must change our trajectory/navigation of this scene (Julian, Ryan, & Epstein, 2017). They explain the relevance of real-world object size in terms of its importance for spatial navigation: since large objects are more likely to be fixed in

space and visually prominent (as opposed to small objects), they make better landmarks. The PPA (parahippocampal place area), amongst other brain regions previously associated with processing navigationally relevant stimuli (scenes, buildings), supports this, as it responds more strongly to large objects than to small objects (Julian et al., 2017).

Studies have shown that medial areas of the occipitotemporal cortex are more responsive to large objects (such as scenes, buildings and landmarks) whereas lateral areas respond more to faces, bodies, animals, and artifacts (Konkle & Oliva, 2012). While preferential activation of certain areas of the occipitotemporal cortex was already shown for special object classes such as scenes versus faces, Konkle and Oliva (2012) have shown that this organization structure is also applied to regular everyday objects. They have shown that large objects, such as vehicles or furniture, activate areas that are more medial in the ventral occipitotemporal cortex and that small objects, such as mugs or coins, activate areas that are more lateral.

Based on results from brain imaging studies, object sizes differ in where they are stored and how they are processed in the human brain. Any attempt to simulate human processing of object size should therefore differentiate between small and large objects. I have divided objects into different size categories based on estimated averages for prototypical representations of these objects and the spherical bounds enclosing them. Figure 22 shows some examples of these estimated dimensions in inches.

door	36x72	
cabinet	32x34	
closet	83x72	
chair	28x37	
box	24x24	
window	26x37	
clock	10x10	
white_board	48x36	

Figure 22: Estimated object dimensions in inches

I have assembled a collection of free 3D models from different sources, each with vastly different scales. Although these size estimations were made in inches, 3D object models are created in arbitrary units. Objects were therefore manually assigned to size categories. A subset of objects and their assigned size category is shown in Table 3.

Object	Size Category
boots	XS
back_pack	S
desk	M
closet	L
wall	XL

Table 3: Sample objects and respective size categories

Each size category was scaled by a specific factor as can be seen in Table 4 (although there was some variability within this). For example, a chair and a laptop are clearly from two distinct sizing categories. Within their appropriate sizing categories however, the size of the laptop and the chair should be variable. Chairs can take on many sizes and forms, but they are still “chair-sized”. Complications can arise if considering dollhouse sized chairs, chairs for toddlers, or ridiculously oversized chairs. For this thesis, only typical representations of objects were considered, although it is important to consider these other options when estimating average size as well as for creating a more representative model.

Size Category	Vector3 Scaling Factor
XS	(1, 1, 1)
S	(2, 2, 2)
M	(3, 3, 3)
L	(4, 4, 4)
XL	(5, 5, 5)

Table 4: Size Categories and Scaling Factor

4 Results

There were 30 before and after pairings of scenes produced in Unity (60 total images). The images can be found in Appendix A. Casual observation suggests that all 30 after scenes are better than the before image for both placement and scaling. While all the after scenes were substantially better than before, it should be noted that two scenes (Scene 8 and Scene 25) were only moderately improved; while the placement and scaling of the objects were better, the overall scene was still odd.

Although the after scenes are obviously better than the before scenes, an objective measure of improvement would be preferable. There are two ways to do this. First, one might come up with an automatic measure of (psychological) realism in the scenes. This would require an additional database of how objects should be placed, because using the same database that was used to generate the images would result in inevitably high score. However we do not have another database.

The other way would be to have human raters judge how good the images are. This kind of evaluation would be a promising area for future work, however time restraints prevented an empirical assessment for this project.

4.1 Evaluating the output

This section will evaluate the new model by contrasting previous SOILIE 3D output with output created by applying the proposed improvements.

While subjectively evaluating my own scenes is clearly biased, the intention of this thesis was to create scenes that were

more likely to resemble something imagined by a human. At first glance, one can make quick inferences about objects and their relations (what is fixed and what can move or fall over) as well as observe key properties (size and clutter) and create a causal story to better understand the scene before them (Lake et al., 2017; Park, Konkle, & Oliva, 2015; Battaglia et al. 2013). Therefore, I wanted to evaluate the success with only a quick glance; without much scrutiny, which image was clearly better? Scenes were evaluated in terms of object scale and object positioning and as a whole believable (in terms of human imagination) scene.

Although all the after images of the scenes were an improvement of the before images, there was still a variation in regard to the degree of success achieved: some were fully believable scenes while others still felt unnatural. Figure 23 shows an example of a before/after pairing that is not only improved, but also fully believable as a scene a human could imagine. The objects are scaled accordingly and in appropriate places with respect to each other to create an image of believable room. These few objects, with their appropriate sizes and location are easily translated into an idea of a bedroom. It fits with a prototypical idea that we have of a bedroom (e.g., bed, desk, and some accessories). It could be a dorm room, or a bedroom in an apartment or a childhood home. The scene feels natural and it is easy to see this and imagine a story to go with it.

Figure 23: Example of a believable scene

Some scenes were still odd, although scaling and positioning of objects was still an improvement from the before scene. While a big factor in this oddness is the selection of objects, there are also issues with the placement of certain objects. An important issue is that walls do not always align with windows, plaques, blinds, etc. Windows were sometimes placed too low on the ground. If taking into account bay windows or egress windows, then being set at this position would be perfectly reasonable. However, due to this being the only window model (and this type of window is typically set at least 3 feet above the floor or slightly higher than a table) it is not reasonable. Another issue is that scenes only have one or two walls, and do not always have a floor, and end up feeling incomplete. Figure 24 shows the before and after images of the two moderately improved scenes.

Scene 25 in Figure 24 feels unnatural. While we can see that the objects are placed better than in the before image, and sized appropriately, the scene still feels odd. It is strange to have a window so low to the ground and for kitchen counters to not be against a wall. The bookcase seems to not fit in this scene either. Although object selection was not part of the scope of this thesis, it still makes for an awkward scene, especially placed where it is.

The other scene in this Figure, Scene 8 is also a moderate success. Objects are sized better, and have slightly better positioning. However, it seems unlikely that given these objects and asked to form a scene that a human would choose to do so. Having the walls parallel like this and with objects placed on both sides of the wall, it feels like delimiting multiple scenes or rooms and does not feel very cohesive.

(i) Scene 8: before

(ii) Scene 8: after

(iii) Scene 25: before

(iv) Scene 25: after

Figure 24: Moderately improved scenes

5 Discussion and conclusion

This section explores SOILIE 3D's approximation of the human imagination by exploring the differences between artificial intelligence and the human mind, considering the limitations by which it is governed, reviewing other artificial intelligence models of imagination, and possible future directions to be considered.

5.1 General discussion

In order to build systems capable of creating believable 3D scenes, we must first understand how scenes are created in human imagination. Although there is already quite a bit of research on the uses of imagination, there is a gap involving how exactly mental images are created (Davies et al., 2011). There is also a lot of research on how humans understand scenes, with a notable lack of research on how humans use their imagination to create scenes.

If we look at imagination in the sense of mental imagery, it can be broken down into two main stages: determining what objects need to be in an image and creating a mental image of those things (Davies et al., 2011). My work assumes that this first stage of imagination was completed successfully and focuses on improving the second stage: image generation. More specifically, the objective of this thesis was to illustrate the key aspects of human scene creation by procedurally generating scenes in a virtual 3D environment (Unity) using scene descriptions generated by SOILIE 3D.

As previously mentioned, imagination is partially composed of our beliefs, memories, and our perception of the world. On the other hand, SOILIE 3D calculations of object placement are partially based on the ‘memories’ from previous scenes (SUN 3D data) but are lacking any so-called ‘beliefs’ of the world. Mathematical calculations alone are not enough to simulate human imagination; a belief system needs to be implemented to simulate constraints that our world knowledge imposes (gravity, degrees of freedom, etc.).

5.1.1 AI vs the human mind

This section will explore some possible explanations for why the images generated by SOILIE 3D differ from human imaginings of scenes.

Common sense

While most humans might be adept at inferences based on quick scene snapshots, it is hard to encode this common-sense knowledge base supporting these rapid conclusions in a computer-readable way. Although artificial intelligence has achieved performance that equals that of humans in certain aspects, there are also key aspects at which it fails (Lake, Ullman, Tenenbaum, & Gershman, 2017). For example, in their 2017 paper, Lake et al., included images with captions generated by a deep neural network. Despite getting the key object of the scene correct, one could say that the deep neural network completely "missed the point" of the image; while it might get individual components correct, it often fails at seeing the big picture. While a human looking at the image could clearly see scared people trying to escape the wreckage of a major flooding, the caption generated was something along the lines of “people standing on a beach”. Another image was captioned as “woman riding a horse on the dirt road”, overlooking the fact that a person was being dragged behind the horse. In a flash we can see the jarring difference between the image and caption presented. Upon further inspection, we can see that the caption is less wrong than initially thought and closer to being factually correct, albeit missing the point. While deep learning

has moved beyond simple object recognition and towards scene understanding, current algorithms are still better at the former and often get causal relationships between key objects wrong in the latter (Lake et al., 2017).

While humans can be thought of as natural pattern-finding machines, a procedural generation algorithm can be thought of as a pattern-creating machine (Short & Adams, 2017). Procedural generation can be used to model reality and has been used in video games and movies to quickly generate lots of objects or effects. Civilizations, graphical effects, terrain, weather, and responsive narrative are just some of the examples that are often modelled with procedural generation (Short & Adams, 2017).

Procedural generation can create infinite and unique experiences and an essential part of managing the output necessitates a fine balance between boredom and chaos (Short & Adams, 2017). An infinite amount of possible outcomes might seem ideal, however this is not always a good thing. Short and Adams (2017) stated that, regardless of the design of a game, sheer variety is not enough for a player to enjoy it and the same can be said for scene creation. The trouble with inhuman creativity and infinitely unique outputs is that these outputs can be very strange, and lead to implausible scenarios no human would ever think of. As evidenced by past SOILIE 3D output, just because something is possible, does not mean that it is interesting or worth creating. A hundred iterations of a scene containing nothing but the same 5 objects and minute differences would eventually become redundant and unlikely to be of interest to a human.

While artificial intelligence is advancing at an admirable pace, it is easy to take its “intelligence” for granted and be shocked at its shortcomings. Since computers do not ‘think’ about things the same way that humans do, they occasionally produce logical extremes which are completely beyond the reaches of human imagination (Short & Adams, 2017). The current SOILIE 3D output is a good example of this. While some errors are more obvious (e.g., gravity), it is not always obvious why a scene is ‘wrong’ or how it can be improved upon. I believe that placing certain restrictions to limit the ‘randomness’ and create a sense of a ‘belief’ system or ‘common-sense’ will bring SOILIE 3D’s output to a more ‘human’ standard of scene imagination.

Images from text

When it comes to text comprehension, mental images appear like a model of the situation (visuo-spatial nature) which is constructed parallel to the actual representation of the text (propositional nature) (Arnold & Lebrun, 1992). These visuospatial representations also include extra-linguistic context and world knowledge. The representation of an image is essentially of a quantitative order (geometrical and algebraical) whereas the representation of the verbal statement is essentially of a qualitative order (spatial and non spatial) (Arnold & Lebrun, 1992).

The automatic transcription of a verbal statement into an image (2D, 3D, fixed, or animated) is a subject at the intersection of a number of fields in artificial intelligence or similar to it (assisted animation, automatic comprehension of an image, computer aided design and manufacturing, robotics, and human-computer interaction) (Arnold & Lebrun, 1992). One of the main problems with transcribing a verbal statement into an image is that the informational content of these two types of

representations is different (sequential vs spatial) and the manner of processing this knowledge differs as well. Human use of perceptual procedures (such as identifying symmetries, angles, distinguishing features from backdrops, etc.) should play a fundamental role in articulating the spatial and sequential representations in order to solve this problem (Arnold & Lebrun, 1992).

The main challenge is in reducing the gap between these representations and finding procedures to pass from one to the other. Language and images have complementary roles and are inextricable from each other. Language permits us to formulate what is not visible in the visual realm; it can describe the functional aspect of morphology, the action of a human in a physical environment, or simulate qualitatively static or dynamic situations (Arnold & Lebrun, 1992). Language also provides us with a way to reason at different levels of abstraction. Visuo-spatial representations of an object or a situation facilitates the integration and the improvement of information provided by the verbal statement on it (Arnold & Lebrun, 1992). It also supports inferences; a stereotypical scene triggers a stereotypical event that is attached to it. For example, if we know that someone is relaxing at the beach, we can infer that they are not working at that moment (Arnold & Lebrun, 1992).

Mental simulation

Mental simulation is an important feature of human cognition and is defined as the ability to imagine what could be (Hamrick, 2019). While mental simulation is useful in scene understanding, I believe it can also help inform scene creation. In recent years, artificial intelligence has been developed to simulate mental models and mental simulation and, although these model-based methods in deep learning have been very important to building and scaling cognitive models, there is still a large divide separating these models from human mental simulation in terms of generalization, efficiency, compositionality, and creativity (Hamrick, 2019).

Let us imagine a cabinetmaker's workshop. Workbenches that are littered with screws, drills, clamps, and off-cuts piled two feet high. A woodpile in disarray, with sawdust and wood shavings swept into random piles along the shop floor and garbage bins scattered by machines. Dust collection tubing that is creating a maze along the floor. Not to mention the many machines that are running and have people feeding long pieces of wood through them. Lake et al. (2017) would suggest that in order to successfully navigate and perform tasks in this chaotic environment, an autonomous agent would need intuitive physics to correctly reason about stability and support. Additionally, a holistic model of the scene would necessitate the composition of individual object models and their relations. They also state that causality is necessary in order to connect different object models correctly, since it supplements the recognition of tools (or learning of new ones) with an understanding of their use (Lake et al., 2017). For example, hand planing the edge of a piece of wood clamped to the side of a workbench, or an outfeed table catching sheets of melamine being cut on the table saw. It is next to impossible to truly understand the actions of the workers in this workshop scene without considering their goals and intentions towards each other and other objects. Lake et al. (2017) attribute the shortcomings of artificial intelligence at perceiving nuances in scenes to its lack of intuitive

physics, intuitive psychology, compositionality, and causality.

Intuitive psychology, also known as folk psychology, is for inferring social causality and is comprised of low-level social perception and higher-level social intelligence (Baron-Cohen, Wheelwright, Spong, Scahill, & Lawson, 2001). While it is important for understanding scenes, my focus was on placing inanimate objects into a scene rather than including live agents, so it was irrelevant in this application.

Compositionality

Compositionality is the idea that primitive elements can be combined in order to create new representations (Lake et al., 2017). This has greatly influenced cognitive science and artificial intelligence, especially in theories of object recognition, conceptual representation, and language. A finite set of primitives can be used to create an infinite number of representations; humans can produce an infinite number of sentences, the mind can produce an infinite number of thoughts (Lake et al., 2017). In programming, primitive functions are combined to create new functions which in turn can be combined to create even more complex functions. Deep neural networks have a limited notion of compositionality that, if further developed, can lead to faster and more flexible learning. In order to properly replicate the human mind's compositionality, a model must maintain the notion of coherence when understanding novel configurations as well as explicit representations of objects, identity and relations (Lake et al. 2017).

In theory, SOILIE 3D should be able to create an infinite number of scenes. Issues arise because SOILIE 3D does not have explicit representations of objects and their relations. A `table_top`, `table_leg`, and `table` will be put into a scene and treated as three separate objects rather than the components of one object. Multiple tables are also set in a scene without any chairs and windowsills and blinds are set in scenes without any windows.

Causality

Causal reasoning is important for supplying clear explanations and it is one of the main hallmarks of human imaginative reasoning (Mahadevan, 2018). A desire to understand the world in simple "cause-effect" relationships drives humans to seek causal explanations and this is becoming more important now that autonomous systems are becoming more prevalent (e.g., accidents that involve a self-driving car might require explanations, usually of a causal form, to be provided) (Mahadevan, 2018).

Walker and Gopnik (2013) describe the links between imagination and causal reasoning; causal knowledge allows us to generate counterfactual inferences, which is tied to our ability to imagine possible worlds. They focus on developments in computational theories of human learning and explore this from a probabilistic Bayesian perspective. Knowing about how events are causally related allows us to imagine creating new relationships or altering existing ones. According to a Bayesian account, we form rational inferences based on the probabilities of possibilities. The likelihood of these possibilities are updated as we accumulate more evidence about the underlying causal structure of the world. Learning is therefore represented

by these more informed inferences that are better approximations of the truth (Walker & Gopnik, 2013).

In scene understanding, causal models represent theoretical real-world processes that generate perceptual observations (Lake et al., 2017). Causality has been influential in perception, with “analysis-by-synthesis” theories asserting that modelling the process that generated sensory data allows for a richer representation of it as well as a richer basis for generalizing in new ways and to new tasks. Causal knowledge also influences how humans learn new concepts by changing how they learn and generalize. For example, the feature “flammable” is more associated to wood than money despite it being equally applicable to both (Lake et al., 2017). This is because there are more underlying causal roles between wood and flammable than money and flammable. Chairs are associated with tables, and blinds and windowsills are heavily associated with windows. We expect these objects to go together because we have been exposed to these associations and have come to expect them to go together. SOILIE 3D has no knowledge of object associations and so there is a large disconnect between objects.

Lake et al. (2017) state that people also understand scenes by building causal models. They explain that humans combine intuitive physics, psychology, and compositionality into a story that reaches beyond perceptual observations (causal model). Perception without these key elements and the causality binding them will lead to errors such as those mentioned previously (flood for a beach, horse dragging a person). Lake et al. (2017) also emphasize the importance of causality in infusing an object (familiar or new) with its use and in connecting object models in the correct way.

This lack of causality can explain why SOILIE 3D scenes feel disconnected from reality. There seems to be no relation between the objects in a scene. Although there is nothing technically wrong with placing some objects in a scene together without other objects strongly associated with them (blinds and windows), there is no application of physics and no coherent ‘story’ to be found within the scene which leaves the viewer confused.

5.2 Limitations

This section will describe the limitations of SOILIE 3D; while there were significant improvements to the original output of SOILIE 3D, there is still room for improvement. Due to the limited scope of this thesis, I had to use simplified constraints to simulate ‘common-sense’ judgements that humans make when imagining scenes.

Object rotation

Object rotation did not fall within the scope of this thesis but I did include a simple constraint to solve part of this issue. In the previous version of SOILIE 3D, objects were allowed to rotate freely on all their axes. This meant that lamps were placed on their sides, and tables could be placed upside down. While this is possible, it would be an unlikely choice for a human to make. I therefore manually assigned a bottom-side to the object models and constrained the models to always be upright (bottom-side down).

Object selection

There were limitations to object selection (in scene descriptions and in modelling the images). Scenes look incomplete with only one wall or two, or just a floor. Windows look strange floating in space and blinds just hanging around are also strange. This assortment of objects was chosen based on sufficient triplet data, but these five objects do not always make up cohesive feeling scenes. However, with other objects in the scene, some of these placements would make more sense. As it is, with just these objects, it does not make sense for them to be so spaced apart. It also doesn't make sense in terms of compositionality for these objects to be placed together. Some scenes work great (bed, window, blinds, plaque, wall) while others seem disparate and contrived. I did not attempt to fix this issue as I took this step in the imagination process (selection of objects to fall within a scene) to be complete. For the imaging aspect of this thesis, only one object model was assigned per object label. Not all models are from the same database and some are even missing (replaced with 2D sprites). While impacting the variability of scenes (hugely lacking compared to the vast reserves of human imagination), this also affects the positioning of these objects (as previously explained in section 5.2). Future work could implement more object models that would account for these variations in object position. In SUN 3D database, regardless of the window type, all windows were labelled the same. This could affect the data in the calculation of window placement as there are different standards to the placement of specific types of windows. The same can be said for other types of objects such as plants (big potted plants vs small succulents) or clocks (wall clocks, alarm clock, grandfather clock, etc.).

Object labels

A known problem with machine learning is that there is a need for large amounts of labeled data for training, and this data is scarce and obtaining more is expensive (Stewart & Ermon, 2017). As mentioned previously, object labels were taken from the SUN3D database, and for some objects, SOILIE 3D takes labels of object parts and does not treat them as a whole. Humans are usually able to learn without explicit examples, and prefer high level instructions for the task at hand, or an idea of what it should look like once completed (Stewart & Ermon, 2017). Scene descriptions generated by SOILIE 3D have included object components rather than separate objects. If asked to select five objects with which to populate a scene, humans would choose five whole objects and not treat object components (table top, table leg, table, etc.) as separate objects.

Rather than use direct examples of input-output pairs, Stewart and Ermon (2017) introduced a novel approach of supervising neural networks by specifying constraints derived from prior domain knowledge (e.g., laws of physics) that should hold over the output space. Although their approach brings up new challenges for encoding prior knowledge, they were able to successfully train a "convolutional neural network to detect and track objects without any labeled examples" (Stewart & Ermon, 2017).

Changing the focus from object labels to constraints based on domain knowledge, such as the effects of gravity, would be beneficial to a more realistic approach of modelling human imagination. Although I have begun to implement such constraints,

they are simplified and do not cover all aspects of our knowledge of objects and their relationships.

Properties of scenes

Scenes can be labeled with semantic category descriptors (stadium, pantry) as well as geometrical properties (stadium is big, park is small) (Park et al., 2015). Psychophysical research indicates with just a glance our visual system can absorb incredible amounts of information describing the scene, such as the size of the space, its location (indoor/outdoor), the nature of its boundaries (enclosed/open), its level of clutter (full/empty), its navigability (yes/no), the nature of the objects composing it (natural/manufactured) (Park et al., 2015). Despite much research on scene properties extracted in a glance, there is still much to be understood about how and where the brain represents these properties. Park et al. (2015) explain how objects and scenes in the real world are characterized by a rich taxonomy of categorical structure.

When humans enter a new environment, or view an image of a novel scene, there are two general properties that are immediately accessible and constrain our action or navigation in this environment: size, and functional clutter. Park et al. (2015) define size as the “spatial extent bounded by walls” and clutter as the “organization and quantity of objects that fill up the space”. Park et al. (2015) examined how regions along the occipital-temporal lobe responded to pictures of real-world scenes and found that the dimensions of size and clutter are parametrically coded in the brain, property-based, flexible enough to generalize to different semantic categories, and can help explain our rapid scene recognition and navigation abilities.

They found evidence that size and clutter are represented in the patterns of parahippocampal place area (PPA), with the retrosplenial complex (RSC) activity patterns being more sensitive to size rather than clutter. They also stated that the size and clutter information was presented in a way that generalized across different semantic categories and that the data supports a property-based representation of spaces, distributed across multiple scene-selective regions of the cerebral cortex (Park et al., 2015).

Studies have shown support for object-based representations of scene-category information which extends beyond the classic scene-processing regions (Park et al., 2015). Recently, neuroimaging studies have started examining the hypothesis that scenes are also represented based on global properties which describe the spatial geometry and featural content of the scene independent of objects. “Openness”, for example, is represented in the PPA, while “naturalness” is represented in the lateral occipital cortex (LOC). Due to their similarity in spatial layout and despite being from different semantic categories, a street and a forest show similar patterns of activation in the PPA. The PPA and RSC and not LOC are sensitive to the spatial layout of a scene. The PPA and LOC and not RSC are sensitive to the presence of an object inside a scene. This provides evidence for a property-based representation of scene information (Park et al., 2015).

This can help explain why we disagree more with SOILIE 3D scenes where objects are clustered on top of each other, or in scenes where the objects are so far apart they are barely visible on the screen, rather than the actual objects within the scene.

5.3 Related work

This section will explore research relating to SOILIE 3D and the use of artificial intelligence in imagination and image generation.

When considering object locations and memory, it is important to not focus solely on descriptions of static environments viewed from an external perspective, but to also consider that a human reading a spatial description will probably imagine moving through this environment at some point in time (Deyzac et al., 2006). The processing of landmarks and landmark positions are supported by distinct working memory resources and our memories for route and survey descriptions are supported by distinctive cognitive processes (Deyzac et al., 2006). The concept of location is relative; to specify the location of an object, a frame of reference is needed (McNamara, 2002). The location of a chair can be described in relation to the observer (e.g., the chair is behind me), in relation to other chairs (e.g., the chair is in the last row, two from the left), or in relation to the room (e.g., the chair is in the top right corner of the room). When noting the locations of objects in novel environments, humans rely on a spatial reference system inherent to the objects being used. Axes and directions are chosen based on viewing perspective (e.g., egocentric) and the bounding area of the environment (e.g., walls surrounding a room) (McNamara, 2002). Some researchers believe that spatial memories are mostly composed of object-to-object spatial relations, while self-to-object spatial relations are mostly used in visual memories of layouts (McNamara, 2002; Deyzac et al., 2006).

Machine learning or data science involves constructing a probability distribution from samples or generating novel samples that are close to indistinguishable from given samples (Goodfellow et al., 2014; Mahadevan, 2018). Although most machine learning is based on using statistical correlations from data, human achievements in the arts, science, and literature require more than simply analyzing correlations. Imagination science can extend far beyond the realm of data science by using causal reasoning, by generating samples that are novel, and by using analogical reasoning to understand novel situations (Mahadevan, 2018).

Left unexplored by most AI research, imagination is one of the features of human intelligence and is considered important in a wide range of fields, from the arts to literature and science (Mahadevan, 2018). There is limited research on artificial intelligence and imagination, with most of the research focusing on reproduction by repeated data sets or on the creative aspect of imagination. SOILIE 3D is the first attempt at a cognitive model using procedural generation in order to replicate the human processes used in imagination. I did not use deep learning because there was a lack of available data.

Although there are other programs that use text input to generate images (e.g., DALL-E, WordsEye), they are not intended to replicate human processes and can be psychologically unrealistic.

DALL-E is an artificial intelligence program that can create images (of either realistic or unrealistic objects) by interpreting natural language prompts (Ramesh et al., 2021). It can create images from unusually abstract concepts (e.g., given the input "a tapir made of accordion", it has produced images of an accordion with parts shaped like a tapir's trunk or a tapir that had

an accordion for a body) (Ramesh et al., 2021).

Coyne, Zamchick, and Bauer are the founders of WordsEye, an original application that allows users to "Type a Picture" (Coyne & Sproat, 2001). It is intended as a novel social network for "visual banter" with users being able to share images in the WordsEye Gallery or on existing social networks such as Twitter or Reddit. The user's text input is first broken down into a semantic representation. This necessitates a large database comprised of linguistic and world knowledge of object parts and properties. Using ray tracing, these semantic representations are then rendered into 3D scenes, which can be endlessly modified (Coyne & Sproat, 2001).

Unlike AI programs like those mentioned above, humans have a rather limited working memory. It is quite possible for humans to experience an overload of visuospatial working memory (Bilda & Gero, 2007). Working memory retains information in visuospatial and/or verbal modes and is an ideal workspace for cognitive processes, such as designing a scene. Maintaining and transforming visuospatial information requires central executive resources and imagery activity uses up an extensive amount of working memory resources (Deyzac, Logie, & Denis, 2006; Bilda & Gero, 2007). The capacity of the visuospatial working memory is limited further when visuospatial tasks are done using solely imagery (Bilda & Gero, 2007).

There are many programs that use machine learning frameworks to generate images, however they do not try to replicate human processes. Although Generative Adversarial Networks (GANs) can use samples of a fixed distribution to create images, these images are reproductions of existing art and nothing like the imaginative masterpieces created by man (Mahadevan, 2018). Elgammal, Liu, Elhoseiny, & Mazzone (2017) proposed a new system for generating art based on GANs called a Creative Adversarial Network (CAN), however this architecture is specific to art and currently lacks the generality to provide a computational framework for imagination (Mahadevan, 2018). Other models that build upon GANs (e.g., Cycle-Consistent Adversarial Networks) are currently focused on visual domains and capable of specific generalizations only (e.g., translating special characteristics of one image collection onto another image collection without paired training examples) (Zhu et al., 2017; Mahadevan, 2018).

The ability to label particular objects or scenes is of main focus in many AI tasks, such as computer vision or speech recognition; object recognition is a success of machine learning and, more specifically, deep learning (Mahadevan, 2018). Although deep learning performance on fixed datasets has been steadily improving, and can even exceed human performance (in some cases where large amounts of labeled datasets have been used for narrowly defined tasks), barely perceptible noise has been shown to cause these networks to guess the wrong labels. Humans, on the other hand, are not restricted to performing within narrowly constrained problems and are very competent at understanding the multifaceted aspects of the world. Children learn effectively without the need for much explicit labeling and they can interpret objects far better than AI systems today. They are more fluid learners and have a creative freedom.

We know that objects are more than their descriptive label; they are also perceived in terms of actions they enable agents to

do. These affordances are central to imaginative thinking in perception and are essential to the design of everyday appliances. To develop imagination networks that can interpret images like humans, they would need the ability to both recognize and use affordances and the ability to integrate perceptual information into an agent's goals (Mahadevan, 2018).

Recent research in computer vision focuses on model-free, data-driven approaches which are heavily dependent on learning from past experiences and Battaglia et al. (2013) question the plausibility of a model-free approach since their results suggest that combinations of salient features in their scenes alone are insufficient to account for people's judgements. A plausible, model-free account would need to be flexible enough to handle a wide range of real-world scenes and inferences, while also being compact enough to be learnable from people's finite experience, and require additional control mechanisms to decide which features and judgment strategies are appropriate for each distinct context (Battaglia et al., 2013). It would face difficulties when attempting to explain how people perform novel tasks in unfamiliar scenes or how physical scene understanding could interface with their rich language, reasoning, imagination, and planning faculties (Battaglia et al., 2013). There has been progress in using model-free reinforcement learning with deep neural networks, although this requires large amounts of training data and resulting agents lack the "behavioural flexibility constitutive of general intelligence" (Racanière et al., 2017).

5.4 Future Work

Further research in creating a realistic model of human imagination of scenes could include implementing a belief system, or an extensive bank of world knowledge and common sense. Another approach would be to implement an intuitive physics model and refine the constraints for object placement. Improving the efficiency of SOILIE 3D and obtaining more data would also allow for the placement of more than five objects within a scene. Implementing a model of object scaling that represents magnitudes as distributions of fuzzy sets would be a more accurate representation of human understanding of size and a better model of human psychology. Further research could also benefit from a more objective measure of evaluating scenes, either by calculating an automatic measure of psychological realism or with human raters judging images.

5.5 Conclusion

The goal of the present work was to illustrate the key aspects of human scene creation by improving SOILIE 3D's scene descriptions and procedurally generating plausible 3D scenes in Unity. SOILIE 3D was the first step towards the lab's imagination process, but this model was too theoretical and did not take human cognition into account (physics, compositionality, degrees of freedom, space, etc.). Imagined scenes are bound by the knowledge constraints of the individual imagining it, although there are certain common physical properties that must be respected. For example, whether our interactions in space are real or imagined, the mental representations we form of space are different from the external representations of space in

terms of physics, geometry, or maps (Tversky, 2003).

When trying to understand something, whether it is a picture, text, or event, we must create an interpretation that fits best with the information available to us (Thagard & Verbeurgt, 1998). The most coherent account considering all of the information is the best interpretation. Coherence can be described as the "maximal satisfaction of multiple constraints" (Thagard & Verbeurgt, 1998). Although there is not a lot of research on how humans imagine scenes, I believe that by extending the findings of research on human scene understanding towards scene imagination, I was able to implement a more plausible model of scaling and positioning of objects within a 3D scene. I have added multiple constraints that must be satisfied in order to simulate the restrictions of human scene understanding. I have limited the object labels to be included in a scene description to whole objects only, rather than incomplete components. I created a list of possible attachment points and used it to determine whether intersecting objects could be placed above or below each other. If there were no matches in attachment points, then objects were shifted so that no objects were intersecting. Although I did not focus on the rotation of objects, I restricted all objects to be positioned upright, as this was a minor fix and viewing plants and couches on their sides or upside down was needlessly distracting. I created a list of average sizes for objects and sorted them into approximate size categories to more realistically scale objects. To place objects more realistically, I restricted objects to the ground (or another object's surface) or a vertical plane on the perimeter of the bounding box. I also restricted the maximum distance that objects could be placed from others, both horizontally and vertically.

Contrasting 30 images created solely with scene descriptions from SOILIE 3D with 30 images created with the addition of my modifications, it can be seen that the changes I have made have resulted in images that are more representative of a psychologically plausible model of imagination. Although the exact manner by which these constraints are implemented in the human mind are unknown, the resulting images had better positioning and scaling of objects and made for a more coherent representation of a 3D scene. As shown in the Results section, all of the images generated were more consistent with imagining scenes from a human perspective than the previous version of SOILIE 3D. There are no longer unconventionally sized objects intersecting with other object's components and floating at bizarre distances from the ground. The creation and comparison of these 30 pairs of before and after images can further help inform the important features that humans consider when imagining scenes. Although these results were an improvement from the previous version of SOILIE 3D, there is still much more work to be done in order to truly simulate human imagination.

6 References

- Archambault, A., & Venet, M. (2007). Le développement de l'imagination selon Piaget et Vygotsky: d'un acte spontané à une activité consciente. *Revue des sciences de l'éducation*, 33(1), 5-24.
- Arnold, M., & Lebrun, C. (1992). Utilisation d'une langue pour la création de scènes architecturales en image de synthèse. Expérience et réflexions. *Intellectica*, 15(3), 151-186.
- Baron-Cohen, S., Wheelwright, S., Spong, A., Scahill, V., & Lawson, J. (2001). Studies of theory of mind: Are intuitive physics and intuitive psychology independent. *Journal of Developmental and Learning Disorders*, 5(1), 47-78.
- Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45), 18327-18332.
- Bilda, Z., & Gero, J. S. (2007). The impact of working memory limitations on the design process during conceptualization. *Design Studies*, 28(4), 343-367.
- Breault, V., Ouellet, S., Somers, S., & Davies, J. (2013). SOILIE: A computational model of 2D visual imagination. In *Proceedings of the 11th International Conference on Cognitive Modeling*, Ottawa: Carleton University.
- Bukowski, R. W., & Séquin, C. H. (1995, April). Object associations: a simple and practical approach to virtual 3D manipulation. In *Proceedings of the 1995 symposium on Interactive 3D graphics* (pp. 131-ff).
- Coyne, B., & Sproat, R. (2001, August). WordsEye: An automatic text-to-scene conversion system. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques* (pp. 487-496).
- Davies, J. (2019). *Imagination: The science of your mind's greatest power*. Pegasus Books.
- Davies, J., Atance, C., & Martin-Ordas, G. (2011). A framework and open questions on imagination in adults and children. *Imagination, Cognition and Personality*, 31(1), 143-157.
- Dawes, A. J., Keogh, R., Andrillon, T., & Pearson, J. (2020). A cognitive profile of multi-sensory imagery, memory and dreaming in aphantasia. *Scientific Reports*, 10(1), 1-10.
- Deyzac, E., Logie, R. H., & Denis, M. (2006). Visuospatial working memory and the processing of spatial descriptions. *British Journal of Psychology*, 97(2), 217-243.
- Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (2017). Can: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*.
- Farah, M. J. (1989). The neural basis of mental imagery. *Trends in neurosciences*, 12(10), 395-399.
- Gagné, J., & Davies, J. (2013). Visuo: A model of visuospatial instantiation of quantitative magnitudes. *The Knowledge Engineering Review*, 28(3), 347-366.
- Ganis, G., Thompson, W. L., & Kosslyn, S. M. (2004). Brain areas underlying visual mental imagery and visual perception: an fMRI study. *Cognitive Brain Research*, 20(2), 226-241.

- Goldman, A. (2006). Imagination and simulation in audience responses to fiction. *The architecture of the imagination*, 41-56.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Gosele, M., & Stuerzlinger, W. (1999). Semantic constraints for scene manipulation. In *Proceedings Spring Conference in Computer Graphics' 99*. Budmerice, Slovak Republic.
- Hamrick, J. B. (2019). Analogues of mental simulation and imagination in deep learning. *Current Opinion in Behavioral Sciences*, 29, 8-16.
- Jacobs, C., Schwarzkopf, D. S., & Silvanto, J. (2018). Visual working memory performance in aphantasia. *Cortex*, 105, 61-73.
- Julian, J. B., Ryan, J., & Epstein, R. A. (2017). Coding of object size and object category in human visual cortex. *Cerebral Cortex*, 27(6), 3095-3109.
- Keogh, R., Bergmann, J., & Pearson, J. (2020). Cortical excitability controls the strength of mental imagery. *elife*, 9, e50232.
- Keogh, R., & Pearson, J. (2018). The blind mind: No sensory visual imagery in aphantasia. *Cortex*, 105, 53-60.
- Konkle, T., & Oliva, A. (2012). A real-world size organization of object responses in occipitotemporal cortex. *Neuron*, 74(6), 1114-1124.
- Kosslyn, S. M. (1994). *Image and brain: The resolution of the imagery debate*. MIT press.
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, 40.
- Lalande, A. (1991). *Vocabulaire technique et critique de la philosophie*. Paris : Presses universitaires de France (1re éd. 1926)
- Mahadevan, S. (2018, April). Imagination machines: A new challenge for artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).
- McNamara, T. P. (2002, May). How are the locations of objects in the environment represented in memory?. In *International conference on spatial cognition* (pp. 174-191). Springer, Berlin, Heidelberg.
- Park, S., Konkle, T., & Oliva, A. (2015). Parametric coding of the size and clutter of natural scenes in the human brain. *Cerebral cortex*, 25(7), 1792-1805.
- Pearson, J., & Kosslyn, S. M. (2015). The heterogeneity of mental representation: Ending the imagery debate. *Proceedings of the National Academy of Sciences*, 112(33), 10089-10092.
- Pylyshyn, Z. W. (2002). Mental imagery: In search of a theory. *Behavioral and brain sciences*, 25(2), 157-182.
- Racanière, S., Weber, T., Reichert, D. P., Buesing, L., Guez, A., Rezende, D., ... & Wierstra, D. (2017, December). Imagination-augmented agents for deep reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 5694-5705).

- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021). Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*.
- Short, T., & Adams, T. (Eds.). (2017). *Procedural generation in game design*. CRC Press.
- Smith, G., Salzman, T., & Stuerzlinger, W. (2001). 3D Scene Manipulation with Constraints. In *Virtual and Augmented Architecture (VAA'01)* (pp. 35-46). Springer, London.
- Stewart, R., & Ermon, S. (2017, February). Label-free supervision of neural networks with physics and domain knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.
- Thagard, P., & Verbeurgt, K. (1998). Coherence as constraint satisfaction. *Cognitive Science*, 22(1), 1-24.
- Tversky, B. (2003). Structures of mental spaces: How people think about space. *Environment and behavior*, 35(1), 66-80.
- Walker, C. M., & Gopnik, A. (2013). 22 Causality and Imagination. *The Oxford handbook of the development of imagination*, 342.
- Watzlawick, P. (1988). *L'invention de la réalité. Comment savons-nous ce que nous croyons savoir ? Contributions au constructivisme* (Trad. A.-L. Hacker). Paris : Seuil (1re éd. 1981).
- Xiao, J., Owens, A., & Torralba, A. (2013). Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE international conference on computer vision* (pp. 1625-1632).
- Zeman, A. Z., Dewar, M., & Della Sala, S. (2015). Lives without imagery-Congenital aphantasia.
- Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).

7 Appendix A: 30 pairs of before and after images of scenes

Figure 25: Before and after images of scenes

Figure 25: Before and after images of scenes

(57) Scene 28: before

(58) Scene 28: after

(59) Scene 29: before

(60) Scene 29: after

(61) Scene 30: before

(62) Scene 30: after

Figure 25: Before and after images of scenes