

Adam Grimley

B00691778

COM668

Live Doc

Virtual Space Compiler

Contents

[1. Introduction 3](#_Toc531784720)

[1.1 Project Statement 3](#_Toc531784721)

[1.2 Project Aim 3](#_Toc531784722)

[1.3 Objectives 4](#_Toc531784723)

[2. Literature Review 5](#_Toc531784724)

[2.1 Part one – Problem domain 5](#_Toc531784725)

[2.1.1 RCNN 5](#_Toc531784726)

[2.1.2 Pose Detection 6](#_Toc531784727)

[2.1.3 Physics engine 7](#_Toc531784728)

[2.2 Part two – Similar Solutions 8](#_Toc531784729)

[3. Project plan and Requirements Specification 10](#_Toc531784730)

[3.1 Stakeholder Identification 10](#_Toc531784731)

[3.2 Requirements gathering methodology 10](#_Toc531784732)

[3.3 Requirements prioritisation strategy used 11](#_Toc531784733)

[3.4 Initial Requirements 11](#_Toc531784734)

[3.5 Proposed System Architecture 12](#_Toc531784735)

[3.6 Software Lifecycle Methodology 12](#_Toc531784736)

[3.7 Work Breakdown Structure 15](#_Toc531784737)

[3.8 Gantt Chart 16](#_Toc531784738)

[3.9 Resources identification 17](#_Toc531784739)

[3.10 Verification plan 17](#_Toc531784740)

[3.11 Validation Plan 18](#_Toc531784741)

[4. Risk Assessment 19](#_Toc531784742)

[5. Initial Prototype 21](#_Toc531784743)

[5.1 Risk and rationale 21](#_Toc531784744)

[5.2 Design Artefacts 21](#_Toc531784745)

[5.2.1 JSON structure 21](#_Toc531784746)

[5.2.2 Training images 22](#_Toc531784747)

[5.2.3 Prototype dataflow 24](#_Toc531784748)

[5.3 Issues in prototype development 24](#_Toc531784749)

[6. References 25](#_Toc531784750)

# 1. Introduction

## 1.1 Project Statement

Through the author’s industrial placement, he had worked on immersive technologies. One major proponent of immersive development is the generation of accurately scaled spaces to preserve special awareness for the user. Using game development as a guideline, we can see that development of 3D environments is often one of the largest costs to development [1].

Many programmers - like the author of this report - have been looking at cutting edge research papers and technologies to see if we can shorten this process down in terms of both time and money. It is important to mention that although the author focuses on immersive development, this project is not just an immersive opportunity however it was immersive technology that lead the way to this research so will be mentioned throughout.

Alongside the author’s immersive work during industrial placement, he had investigated new ML techniques such as regional convolutional neural networks (RCNN) and pose detection. These techniques have recently presented us (the immersive development community) with a possible opportunity to build a pipeline that will allow us to quickly generate 3D spaces based on a flat photograph of a real space. The components of this pipeline will be discussed and developed throughout the course of this report.

An area of concern for the author through this project will be the scope he is looking at. Although a lot of research has been completed for each section of the pipeline, no one has so far proved they can work together. This is hard to protect against other than making sure each section tests succeed in a vacuum before integrating them. As well as this, each section of the pipe is very recent developments and could be unstable for use in a production environment. In an effort to mitigate against any issues that may arise because of this, frequent correspondence with the authors of the key papers [2], [3] will be attempted through development.

This project makes use of 2 major research libraries which have been recently released. The first is mask-RCNN which is a neural network project which allows us to quickly develop an RCNN which will segment the picture a user has input. This is the component that will allow the product to recognise objects and to isolate them from the picture. This software library has been developed by Matterport and open sourced under the MIT License allowing open use of the software.

The second major component to be used in the project is the recently published KeypointNet project. This software allows for the detection of object pose within the user’s image. This section will be placed after mask-RCNN to reduce any error within the detection of each object’s relative pose. This research product was developed by Google AI and released under the apache license. The reason the author has chosen this library is to speed up the addition of future objects to recognise. The KeypointNet shows that it does not need user defined key-points to generate pose models. This will be important in the future because we want to make the pipeline as easy to use as is possible.

## 1.2 Project Aim

The overall aim of this project is to release a way for us to take a single photograph and turn this into an accurate 3D space for use in simulations. The project seeks to, in turn, open more ways for the technology community to hasten the development of accurate 3D locations. The product(s) planned will be an open source pipeline and a research document outlining the end pipeline - assuming success of the project. The pipeline looks to take a photograph and pass this through a trained RCNN and pose detection system hosted in the cloud and return a file which can then be used by the Unity engine to generate a 3D virtual space.

## 1.3 Objectives

Objectives of the project are outlined in figure 1 below. These objectives will be the general aims of the project and outline what needs to be accomplished at a high level before the author will deem the project a success. Each objective was developed by the author to either test the implementations of cutting-edge machine learning algorithms or test his own knowledge of the subject area.

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Objective | Achieved S1 | Achieved S1 |
| 01 | Input a picture from any camera source into system. The initial project scope will use a png image. | N | - |
| 02 | UI built within the unity engine | N | - |
| 03 | Single object recognition from picture | Y | - |
| 04 | Isolate the desired object from a test image | Y | - |
| 05 | Recognise the depth of object in relation to the camera | N | - |
| 06 | Recognise the global rotation of desired object | N | - |
| 07 | Output position, rotation, and depth of object in relation to camera | P | - |
| 08 | Use a JSON structure for data transfer between components | Y | - |
| 09 | Build API to allow the application to offload processing of image | P | - |
| 10 | Use Azure/AWS VM with a public IP for server hosting | Y | - |
| 11 | Offload ML (machine learning) model training to the cloud | Y | - |
| 12 | Generate 3D scene from JSON structure output by API | N | - |
| 13 | Create a selection of models to substitute for objects in 3D scene | N | - |
| 14 | Organise system into single application pipeline | N | - |

Figure 1 - Objective listing. Assigned completion into Y (yes), N (no), and P (partial)

# 2. Literature Review

## 2.1 Part one – Problem domain

To complete this project in a way which furthers the AI (artificial intelligence) community in Northern Ireland and push the boundaries of automated image processing, the author of the project has conducted preliminary research into several topics including R-CNN [3]–[5] and pose detection[2], [6]–[8].

### 2.1.1 RCNN

Before getting into the reasoning of using an R-CNN within this pipeline, a brief overview of the research that led the author to the conclusion that an R-CNN would be useful. The R-CNN (Region of interests with convolutional neural network features) structure was proposed as an effort to bridge the gap that sat between object classification and detection [9]. The basic R-CNN structure split the process into 3 stages or modules:

* Extract region proposals
* Compute CNN features
* Classify regions

Using an RCNN we essentially generate *r* number of regions of varying size across the picture input. These regions are category-independent. There are many different ways to generate these regions as objectiveness[10], multiscale combinatorial grouping [11], and object proposals [12].

After the initial paper outlining the use of RoIs and CNNs together [9], the next major upgrade to the R-CNN format came in the form of Fast R-CNN (the background of Fast R-CNN is required for the comparisons in figure 3). This methodology proposed higher efficiency of the R-CNN method by implementing feature sharing and multi-task loss during training among other advancements [5]. Researchers developing this method had shown up to 18.3x increased training speed than R-CNN and SPPnet [13], another computer vision method [5]. Unfortunately, Fast R-CNN uses object proposal as a method for determining region of interests in the given picture (RoI). Due to the nature of this project, the author decided it best to look for a different method for determining the RoI rather than manually specifying which object to look for each time the application was executed.

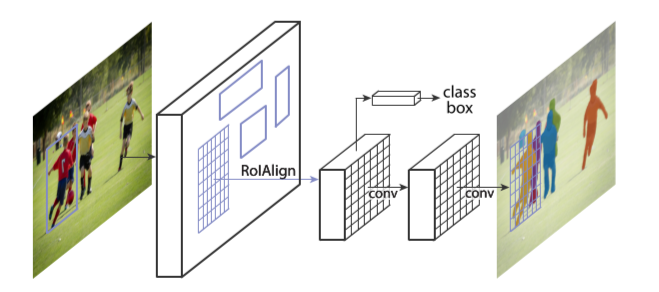
Another route to follow when aiming to fill the object recognition, classification, and isolation was to use the FCN (Fully convolutional network) methodology [14]. Although this is not an R-CNN, it could in theory complete the work required for the object recognition and classification, and image segmentation component of the proposed product. The author’s reasoning against using FCN was due to the work required to adapt the methodology to the current program. Although in terms of standalone semantic segmentation FCN could be a great contender, the product theorised by the author would require the instancing of each class. The FCN methodology alone shows to work poorly for this [3] and would require further work to allow this. In the future, once the pipeline is built, there may be a case for transitioning the component into an FCN if experiments show a speed increase.

Figure 2 - Mask R-CNN framework [9]

The theorised product would require instance segmentation as previously mentioned. Recently work was published detailing mask R-CNN. This is a variation of an R-CNN with a focus on instance segmentation [3]. This is an extension of the previously mentioned Fast R-CNN and as such is fast to implement and train. This relives time pressures on the Author to implement this component.

In terms of the object recognition component of the product, the system would require the recognition of an object within the input photograph. Once an object is recognised, we would then require the separation of this object from the remainder of the picture on a pixel-by-pixel basis. This combines both semantic segmentation and instance segmentation. The only CNN purpose-built to achieve this task in an efficient manner is the mask R-CNN. Figure 3 below shows mask R-CNN currently outperforms all previously mentioned models. The backbone section of the table outlines the neural network architecture of the model. The backbone is worth mentioning when discussing the accuracy of the bounding boxes because it is the net that learns object patterns.

|  |  |  |
| --- | --- | --- |
|  | Backbone | Bounding Box AP |
| Faster R-CNN+++ | ResNet-101-C4 | 55.7 |
| Faster R-CNN w FPN | ResNet-101-FPN | 59.1 |
| Faster R-CNN w G-RMI | Inception-ResNet-v2 | 55.5 |
| Faster R-CNN w TDM | Inception-ResNet-v2-TDM | 57.7 |
| Faster R-CNN, RoIAlign | ResNet-101-FPN | 59.6 |
| Mask R-CNN | ResNet-101-FPN | 60.3 |
| Mask R-CNN | ResNeXt-101-FPN | 62.3 |

Figure 3 - Comparison between R-CNN models in regard to bounding box accuracy percentile on single models [9]

### 2.1.2 Pose Detection

In order for the final product to work as expected, we need to not only figure out what the object is with regard to classification (as discussed above) but we also need 3 pieces of contextual information: position, rotation, and depth. In order to achieve this, a pose detection algorithm will need to be put in place. The main issue with pose research is that many of the papers found by the author focused on algorithms to determine human positions where the product would require object poses.

As context for this component regarding the final product, I am only defining the scope to be a selection of two test images. The images will be comprised of 2 different arrangements of 3 objects on a desk. Each object will have a different pose in the images. As the product is focused on objects, plane detection will not be required for this version of the product and as such, a simple plane will be placed at the base of the lowest object detected.

Modern pose detection processes were first presented in 2005 as pictorial structures for object recognition [7], which is still considered a starting point when developing pose recognition systems [15]. The pictorial structures specify a model that will distinguish between instances of an object based on training models. Again, this paper uses human pose to test and validate the algorithms, but it is worth a mention here as although the author does not require human pose, the graphical tree based approach given by the paper may be required knowledge when developing the object pose model.

From the pictorial structures, we can follow pose detection through hierarchical models [6], non-tree models [8], [16], to convolutional models [15] which are commonplace today. The author wishes to use a pose model which allows us to define which way an object sits in the image. For instance: what way the handle of a cup faces as it sits on a table. The convolutional pose detection model would be able to achieve this however there are several downsides the author had found.

The first of these downsides is the keypoint locations. Specifically, the denotation and training of these locations. In the previously mentioned papers, the methods outlined require the manual specification of each key point for each training image. The author believed that there is a better way to undertake training using other software which will automatically find the key points for itself. This would allow for faster training of models and therefore faster addition of more objects to the system. Shortening training time later in development was seen to be a worthy decision even considering the risk of the required upfront development time [Section 4].

The second downside of these models is the output of point depth in relation to the camera. One major issue when developing a system such as the product proposed, is that depth is notoriously hard to gather from a single point perspective for several reasons. Depth cameras often make use of either infrared or stereoscopic cameras to define depth as each of these technologies allows us to make easy distinctions between locations of objects in relation to the camera. The author however, has specified that the product uses single monoscopic images to create the scene which complicates the process. Rather than taking the depth data direct from the camera, we need to instead analyse the context of the object [17].

In relation to the high-level system architecture the author has proposed in section 3.5 of this report, we need to pass the pose components just the masked image theoretically. This proposal may require changing down the line as with a masked object image we remove the global context. We get around this by using recent research into latent key point analysis. This research proves that using a trained model, we can output the depth of key points within the object image [2].

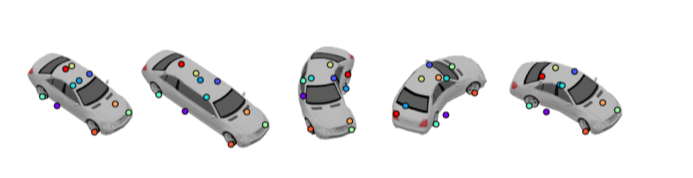


Figure 4 - Keypoints on a deformed car [16]

The main thing that drew the author to this model for depth and pose is the automatic generation of key points. This means that as the system trains, it will define its own key points to use meaning it will always use the most efficient points. The system also analyses and outputs the depth of each key point which can be theoretically used to position the object in relation to the camera. Using this system will reduce the training required for adding objects to the system however does come with a large upfront development cost due to it being experimental technology.

One major risk to using this technology for the process is that it has not been wholly proven against real world scenarios. Throughout the paper [2] the authors have used 3D models to train, test, and validate the model. This may require the author of this project to edit the algorithms used or generate new processes for pose and depth of objects.

### 2.1.3 Physics engine

To complete this project fully, we will need to have a way to develop an accurate way to simulate realistic physics for our virtual space. As the target for this project is for accurate simulations in virtual reality, this is a key factor in completion of the project to the authors specifications. In researching the physics engine to use, a key factor was the licence type and whether the base software is open source or not.

For this implementation of the proposed product, the stakeholders concluded that a fully accurate simulation is not required, and it will be sufficient to have simply plausible simulation physics. This was a key factor in determining the engine that would be used as an initial development point. It allowed the author to choose an engine in which he could be more comfortable working in without needing too many tweaks.

Through research, it was found that in an accurate simulation, the PhysX engine developed by Nvidia was the most stable open source engine although most engines have their pros and cons [18]. The PhysX engine is also the most accessible for the implantation planned as it is the physics backend for the two major simulation development engines the author has experience using: Unity3D and Unreal. The author decided to use Unity3D for this project due to more experience in advanced simulation development on the platform.

## 2.2 Part two – Similar Solutions

There are some similar solutions to the problem outlined. In terms of the transferring of a real life space into a 3D scene that can be used in 1:1 scale would be photogrammetry. Photogrammetry is defined by [19] as:

“the science of obtaining reliable information about the properties of surfaces and objects without physical contact with the objects, and of measuring and interpreting this information.”

When used in the real world, photogrammetry and LIDAR scanning allows engineers and developers to effectively and realistically scan a given real life area in moments. Matterport, One of the global leaders in photogrammetry products states that their scanners can produce a realistic scan of a 2000ft2  area in under 26 minutes [20]. This does not include processing time nor setup time, which increases the total time significantly. These numbers should be taken as anecdotal however, as dealing in scanning the time of the scan depends wildly based on what you need from the location and how many scans you require.

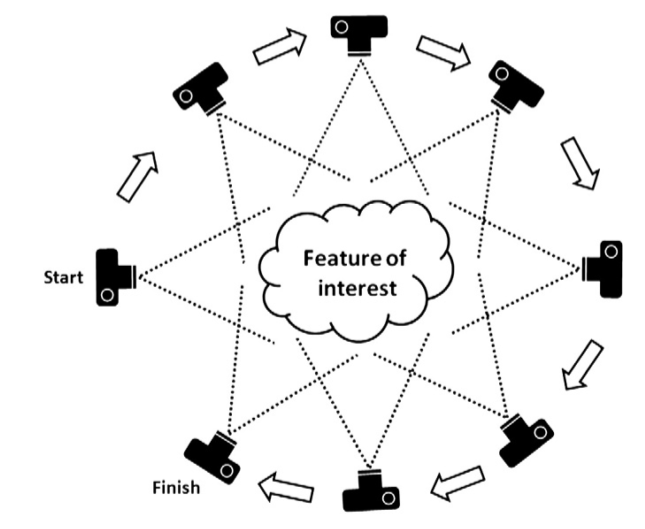


Figure 5 - SfM photogrammetry [21]

The Leica system for LIDAR scanning of an area states a scan time of 3 minutes per single scan [22]. The author has personal knowledge of using this system and understands that a 200ft2 area from start to usable 3D image can take up to half a day. This scan time – although relatively fast for the actions – is still too long for quick developments and each comes with issues in changing the 3D scan to suit a developer’s requirements.

A big difference in the proposed system would be the use of custom models in substitution of real-world textured models. Where a 3D area would be generated based on the textures and measurements from the real world, the proposed system removes the real-world textures. Our system uses the measurements and locations of objects and theoretically, would generate the area faster because of this.

The systems do however have different applications. A geospatial scan such as a scan from LIDAR or generated using photogrammetry have their uses is realistic representations of areas. This is a use case that is key to preserving history and spaces for the future. The proposed system of this report would be used to generate quick 3D spaces using custom models based off a single image. There may be a use case down the line where, using feature recognition or location mapping we could scan and generate full 3D spaces. At the present, due to the research nature of the project, there is only the scope for a single image to be used meaning a restriction in the space generated.

A system such as the product proposed would be used to prototype 3D systems in a realistic space quickly and efficiently. The author came to the idea of this project from developing immersive systems and not being able to test in a real-world scenario without weeks of work beforehand. The system could also be used as a soft segmentation tool to ‘touch-up’ any 3D scans with professionally developed 3D assets.

In researching the applications that are similar to the proposed project, the author had found a recently released paper documenting 3D-RCNN, a convolutional neural network with the purpose of distinguishing 3D features from a 2D image [4]. This works in a similar way to the proposed system however the end product is to be used for different cases. The system proposed by [7] is a real-time system for generating a 3D shape set that can be used for depth and analysis. This is key for autonomous vehicles and other systems that require accurate and fast depth images from 2D graphics.

The system proposed in this report uses object recognition to substitute specified objects in a 2D image with 3D assets in a virtual space. This means that developers may use custom objects to quickly create realistic 3D spaces based off real-world scenarios. Developers using this system are not locked to recreate the real scenario completely.

It is a testament to the research nature of the proposed project that the 3D-RCNN work is not yet published. The paper published is an outline of the algorithms used and how it is built but there is not functioning system so far. Although the use cases are different, the area of work is the same between the author’s proposed system and 3D-RCNN.

# 3. Project plan and Requirements Specification

This section outlines the initial project planning stages as well as justification for requirements gathering methodology and initial specification. As the product is a work in progress, some of the initial requirements may change through the product lifecycle.

## 3.1 Stakeholder Identification

This project has 2 distinct categories of stakeholders: direct and indirect. The direct stakeholders are those closest to the project where the indirect are possible end users. The proposed project, however, is research focused and as such the main goal of the project is to further the development of 2D image recognition technology using recent advances in the field of machine learning. It is due to this, that the main stakeholder listed is the author of this report. All other stakeholders are indirect. These are listed in figure 6 below.

|  |  |  |
| --- | --- | --- |
| Stakeholder ID | Name | Role |
| D01 | Adam Grimley | Developer / Author |
| I01 | Dr Gaye Lightbody | Mentor / Advisor |
| I02 | Kainos Innovation Team | Advisors |
| I03 | PSG | Peer support / Focus group |

Figure 6 - Stakeholder listing

If required by the author, specialist machine learning developers (as outlined in section 1.1 of this report) may be consulted. The author does not intend for this to be required however if they are consulted, they will be added as an indirect stakeholder in a consultant role.

## 3.2 Requirements gathering methodology

In gathering requirements, a meeting was placed with members of the Kainos Applied Innovation team to discuss what would be required of a system like that which has been proposed. The team are recognised members of the Northern Ireland research and development community and as such, have experience in projects such as this. The author conducted this meeting using carefully selected questions designed to create a discussion around the proposed application and to identify and analyse any unforeseen “edge cases” that may come from standard use of the eventual product.

Going forward in the development of the application and research, the author has gathered a focus group to further develop requirements and to receive feedback about the product. This has shown to be a non-discriminatory way of increasing productive input during the development of a product [23] and has been chosen by the author for this reason. These meetings will give people a safe environment to raise concerns about the product direction and to suggest ways in which the product can be improved.

From these focus group meetings, the suggested changes will be taken on board, checked to see if they can fit into the project scope, and passed through as a change request. These requests are examined by the author and checked against the schedule of development to see if they are valuable and time efficient. These may also be discussed with the author’s mentor to examine whether the changes will add a tangible increase in the project viability. Each requirement listed above has been taken from the initial meetings with Kainos and the focus group and may change or be added to as the project is in development.

## 3.3 Requirements prioritisation strategy used

After being introduced to the technique in university, the author decided the best way to prioritise requirements would be to use Karl Wiegers Relative Weighting, evaluating each requirement based on cost, value and risk [24]. In the case of the proposed project, the cost value will represent the estimated time for each requirement to be fulfilled. The main component of the prioritisation strategy is the equation:

**Value percentage / (cost percentage \* cost weight) + (risk percentage \* risk weight)**

This equation allows us to prioritise each requirement based on weights of how important the cost or the risk is. Due to the high risk of this project, the risk weight is set slightly higher than the cost as the author believes time to be less of a factor than the technical complexity. An example of the relative weighting used is in figure 6 below.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Relative Weights: | 1.0 | 1.0 |  |  | 1.0 |  | 1.2 |  |  |
|  | **Relative Importance** | **Relative Penalty** | **Total Value** | **Value %** | **Relative Cost** | **Cost %** | **Relative Risk** | **Risk %** | **Priority** |
| Totals | 122 | 98 | 220 | 100 | 86 | 100.000 | 75 | 100.000 |  |
| Application is standalone | 9 | 6 | 15 | 6.8 | 2 | 2.326 | 2 | 2.667 | 1.234 |
| Application takes single photograph as input | 9 | 8 | 17 | 7.7 | 2 | 2.326 | 1 | 1.333 | 1.968 |
| user does not process the image locally | 7 | 5 | 12 | 5.5 | 7 | 8.140 | 8 | 10.667 | 0.260 |

Figure 7 - Requirements prioritisation example

This process shows which requirement should be handled first. In this case, it would be the single input functionality.

## 3.4 Initial Requirements

Requirements listed in the table below are split into FR (functional requirement) and NFR (non-functional requirement) IDs and linked to the objectives listed in section 1.3 of this report.

|  |  |  |  |
| --- | --- | --- | --- |
| ID | FR/NFR | Requirement | Priority |
| 01 | FR | Application is standalone and does not require a browser | 1.234 |
| 02 | FR | Application takes single photograph as input | 1.968 |
| 03 | FR | User does not process the image locally | 0.260 |
| 04 | FR | Develop a public endpoint server that allows connection to ML models from anywhere | 0.409 |
| 05 | FR | System shall be trained using common objects | 0.451 |
| 06 | FR | System will hold standard 3D models of common objects | 0.536 |
| 07 | FR | System will recognise and report the relative rotation of desired object | 0.450 |
| 08 | FR | System will recognise and report the depth of desired object relative to viewpoint | 0.350 |
| 09 | FR | The system will automatically communicate between different components of the pipeline | 0.352 |
| 10 | FR | The system will substitute 3D models in virtual space based on where the specified objects are in the input image | 0.256 |
| 11 | FR | The user should be able to show/hide metrics during execution | 0.579 |
| 12 | FR | The user should not have to leave the application during the process | 0.829 |
| 13 | FR | The application should have Virtual Reality support | 0.096 |
| 14 | FR | The application should allow the selection of objects to create in the 3D space | 0.680 |
| 15 | NFR | System will be developed using cutting edge open source frameworks for ML | 0.176 |
| 16 | NFR | System will have a way to report metrics including stage times, accuracy percentiles, etc. | 0.536 |
| 17 | NFR | System will not hold any data that isn’t marked as training data | 0.987 |
| 18 | NFR | System shall have a minimum of 70% test coverage | 0.581 |
| 19 | NFR | The API that receives and returns data should be secure | 0.736 |
| 20 | NFR | The application should be desktop only | 1.152 |

Figure 8 - Initial requirements

## 3.5 Proposed System Architecture

As an example of the architecture of the system going forward in planning, the author has outlined below how the system will be structured.

Input

Unity UI

R-CNN

API (Cloud hosted)

JSON Container

Pose Detection

C# Scene Builder

Virtual Scene

3D Model storage system

Output

Figure 9 - Proposed Architecture

## 3.6 Software Lifecycle Methodology

As outlined throughout this report, the full system proposed is comprised of many different components. Each component should in theory be able to work independent from the others however it is the connections between each component that make the system novel. Even so, because of the independent nature of each component, the author has decided that rapid prototyping will be the best practice going forward, so that each component can be prototyped and connected from the ground up in relatively similar timeframes.

It is because of this decision, that the methodology chosen for the project is spiral development. This methodology was first introduced by Boehm in 1988 combining waterfall and RP development [25]. The justification of this is the timeframes of the project set by the university. These given timeframes lend themselves more naturally to a waterfall driven development life cycle which the author believes would not be efficient for a single person development team to implement into this project.

The spiral methodology gives more forgiveness to complex systems [25], such as the system proposed in this report, and allows for faster reaction to unforeseen technical complexity. This is required in the project as some of the technologies that will be used are still experimental and not fully tested. The methodology proceeds as cycles of development, at the end of which there will be a product prototype. At the beginning of the cycle the author will evaluate risks associated and react accordingly – developing the highest risk components first.

In association with the spiral methodology, a Kanban board will also be used. This will be held online and hosted by GitKracken. The Kanban board will be used for tracking the progress of each development cycle. The proposed cycles are listed in figure 10 below. As the project is still in the planning stages, these rounds may vary slightly in the later stages of development in keeping with the methodology.

|  |  |  |  |
| --- | --- | --- | --- |
| Cycle | Task ID | Effort | Description |
| 1 |  |  | **Preparation** |
|  | 01 | L | Proposal |
|  | 02 | H | Initial project plan and report |
|  | 03 | M | Research into component technologies |
|  | 04 | M | Requirements development meetings |
| 2 |  |  | **Basic API** |
|  | 05 | L | Setup cloud VM with public access |
|  | 06 | H | Train RCNN model with single object |
|  | 07 | H | Train Pose model with single object |
|  | 08 | M | Move models into the cloud |
|  | 09 | L | Connect models using JSON container |
| 3 |  |  | **Basic Unity integration** |
|  | 10 | L | Build initial UI |
|  | 11 | L | Integrate the posting of inputs to the VM public IP |
|  | 12 | M | Build initial Scene builder to parse returned JSON |
| 4 |  |  | **Model Building** |
|  | 13 | M | Build initial simple object recognised by API |
|  | 14 | M | Generate structure to hold models |
|  | 15 | M | Integrate structures with scene builder component to spawn objects |
| 5 |  |  | **Additive Development** |
|  | 16+ | M | Train API with additional model |
|  | 17+ | M | Create appropriate 3D model |
|  | 18+ | M | Integrate model with scene builder component |

Figure 10 - Cycles outline

In figure 10 above, cycle 5 will continue to be repeated in the remaining time as in theory all systems will be connected by this cycle and should just need added to. The author only expects one of these cycles to be completed however there is space for continued addition if the time is available.

Each cycle will begin with a meeting with stakeholders to outline and examine any risks associated and will end with all systems tested to assigned standards.

## 3.7 Work Breakdown Structure

Figure 11 - Work breakdown structure

3D space compiler

Cycle 1

Cycle 2

Cycle 3

Cycle 4

Cycle 5

Proposal

Initial plan and report

Research

Requirements Development

Setup VM

Train R-CNN (single model)

Train pose (single model)

Train R-CNN (single model)

Move ML models to server

Initial UI build

Integrate connections to VM

Create initial Scene builder

Create 3D model of object required

Generate model storage structure

Integrate models to scene builder

Train pose (single model)

Create 3D model

Integrate model to scene builder

Create JSON container

Finalise report

Full testing review

Begin final report

Stakeholder review

Application-wide code review

## 3.8 Gantt Chart

Figure 12- Gantt Chart. Tasks in yellow, cycles marked in green. Each task is represented from figure 10.

## 3.9 Resources identification

|  |  |  |
| --- | --- | --- |
| Resource | Type | Purpose |
| Laptop | Hardware | Main form of development. Main piece of hardware for the project and will be used throughout. |
| AWS | Infrastructure | Will be used to host the python API in the cloud. Also will be used to train ML models as a VM cluster will train faster than a standard laptop. |
| Jupyter Notebooks | Software | Python notebook development environment. All python will initially be developed within jupyter as it allows easy and readable modularisation of python files. |
| Git | Software | Will be used for source control throughout the project. |
| Visual Studio 2018 | Software | Used for C# development. Is provided by Microsoft Imagine and will hook directly into the source control. |
| Git Kracken | Software | Online software allowing quick generation of detailed Kanban boards. This will be used to track the cycles of development and any changes that are requested. |
| Unity Engine | Software | Physics engine that will be used to generate all visuals and 3D scenes. Will also call to the API when required. |

Figure 13 - Resources

Due to the project only being developed by a single engineer, human resources have been omitted from identification. This is also because the project is not funded, and these costs are negligible.

## 3.10 Verification plan

The project aims to make use of 2 completely different types of ML models. This adds considerable technical risk as neither have been designed to work together. Each of the specified models will be tested separately using industry standard 80/20 split of the data [26]. This means that each dataset while training the ML models will be split into 80% training data and 20% verification data. Once each model is tested individually, developers will then continue the testing development and start to look at the actual code of the program.

Figure 14 - TDD diagram

Write tests

Pass test

Refactor

Write code

Eliminate redundancy

In testing the inner code of the program, I plan to use an automated workflow. These tests will be broken into unit tests, integration tests, compatibility tests, and user acceptance tests. The use of automation in the tests will provide the developer with the ability to cover all these areas and quickly see if changes break anything already developed and pushed in a previous development cycle.

Throughout the development of both python-based components and C# based components, unit tests will be developed before the full development of the component. This ensures the efficient programming of modules using a TDD methodology.

At the end of each cycle, the current prototype will be put through black box testing. The results of these tests will play a part in defining the work required in the coming work cycle, as outlined in the spiral development methodology [25]. Each black box test will use both the proposed test images as inputs. A verification test for each requirement is listed below.

|  |  |  |
| --- | --- | --- |
| Requirement ID | Verification test | Pass criteria |
| 01 | Run application on a clean install | Application does not require any other software to be installed |
| 02 | Follow application through to image submission | Application only asks for a single image |
| 03 | Submit an image to the system for analysis | Application does not request more resources than are required by a regular unity development session. |
| 04 | Connect to the system from a new IP address | Application allows instant analysis regardless of new IP |
| 05 | When going through image analysis, open object selection window | Application lists several common objects approved by stakeholders |
| 06 | Build a scene using the system | Application automatically uses standard 3D objects with no input required from the user |
| 07 | Analyse an image with the system | Application returns a list of object rotations |
| 08 | Analyse an image with the system | Application returns a list of depth points of the desired object |
| 09 | Run application from start to finish | Application does not require any input once image is submitted |
| 10 | Run application from start to finish | Application generates accurate virtual space based on user input |
| 11 | In the main screen, select to view application metrics | While application is running, it will show metrics on the screen rather than blank loading screen |
| 12 | Run application start to finish | Application stays within a single window throughout |
| 13 | Build a scene using the system | Once scene is built, user can put on a VR headset automatically |
| 14 | On the main screen, open the object selection pane | Objects will be listed |
| 15 | - | - |
| 16 | Run system from start to finish with log option selected | A log folder will be generated on user’s desktop |
| 17 | - | - |
| 18 | Run test coverage program in the backend | Application has 70% or higher coverage |
| 19 | - | - |
| 20 | Try run application on a mobile tablet | Application does not run |

Figure 15 - Verification plan

## 3.11 Validation Plan

The proposed product will be validated through regular meetings with stakeholders. These stakeholder meetings will ideally be held at the end of each cycle and will address any issues that come from the current prototype at the time. All stakeholder will be asked to individually discuss each task outlined in the previous cycle and how the final prototype reflects each. These discussions will produce work items that will be added to the next cycle if the project lead believes there to be enough time for development of these changes before the next major stakeholder review.

# 4. Risk Assessment

The initial risk table is outlined below. This table outlines the criteria of risk, the threat posed by the risk, source of the risk, the control measure to reduce risk likelihood, then the risk scores generated by the author. The final column also outlines the mitigation strategy of the risk. This strategy is the plan of action to take should the risk happen. The risk score is generated using: *Likelihood \* Impact = Risk.*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Criteria | Threat | Source | Control measures | Likelihood | Impact | Risk | Mitigation strategy |
| 01 | Technical | Not enough resources to train models efficiently locally | Local CPU | Use AWS vouchers to move model training to the cloud. Voucher sources either github student or AINI community. | 6 | 4 | 24 | If the vouchers prove not to be enough, contacts from both Kainos and Digital Catapult will be able to supply computers with top of the line GPUs for model training. |
| 02 | Technical | Classification of objects with R-CNN not accurate enough | ML Model training | Spend more time developing a larger training and validation dataset for the R-CNN component | 3 | 7 | 21 | Revisit the training dataset and see if you can improve upon the tagged images. If the dataset is deemed large enough by the project lead, research a different image classification model. |
| 03 | Technical | KeypointNet system does not return the true depth of each object | ML Model | N/A | 6 | 8 | 48 | If the KeypointNet system does not return the true depth, look into monoscopic depth perception algorithms such as [17] and estimate how long it would take to implement. |
| 04 | Technical | VM server could drop due to connection issues | Cloud provider | Using AWS to host the VM server will allow the risk to be almost non-existent. | 1 | 3 | 3 | Host the server locally. |
| 05 | Technical | The JSON data may not translate into appropriate unity scenes | API | This will be mostly a risk with the model data outputs. Care will be taken when training and developing the models to be used such that we return the correct data | 2 | 7 | 14 | Edit the model codebase to allow access to the correct information. |
| 06 | Technical | Unity cannot send the images to the server in the correct format | Physics Engine | The use of a image converter library should allow the translation of an image into a base64 string. This string should be accepted by all formats. | 1 | 2 | 2 | Use a different encoding algorithm on the hosted API |
| 07 | Human | Scope of project may be too large to deliver | Specification / Requirements | The scope of the project should be revisited at the beginning of each sprint and risks mitigated. The scope of the project is large however with continual revaluation, all objectives can be met | 7 | 5 | 35 | The scope should be reduced if it becomes intangible. A one-page document should be produced confirming this and outlining the reasons for the scope reduction. |

Figure 16 - Risk assessment table

To summarise the previous risk table, the highest risks come from the ML models to be used and the overall scope of the project. The author of this report has already deemed the project to be large in scope and understands this risk. This risk is one of the reasons for the spiral life cycle methodology and has been outlined throughout the planning section this report. The technical complexity comes from the use of experimental systems that are being used and tested. These systems are only recently published and have not been proven in production environments as far as the author has researched. Extensive research has been carried out concerning these systems and the author has concluded that – although it is an untested environment – both systems will be able to perform the task within the product with great efficiency.

Medium risks come in the form of data outputs and training datasets. Each of these risks have low likelihood yet high impact on the project should they appear. They are both easily mitigated if they appear, with one only requiring a larger dataset and the other requiring the editing of system codebase. The main impact of these is that they will be time consuming to deal with but not complex (based on preliminary research carried out by the author). They also each have the risk of unknowns. The code of each model used is largely abstracted and may not provide us with the correct data we need. If the training dataset is greater than 100 images total for each, the object chosen may be too generic for the system to correctly distinguish from an image. If this happens, a new dataset will need to be created for a new image which is heavily time consuming and will cause missed deadlines.

The remaining risks have been deemed low risk to the project. These are outlined in green on figure 16. These risks, although they will have an impact, either have an extremely low likelihood or are easily mitigated against.

# 5. Initial Prototype

## 5.1 Risk and rationale

In developing the initial prototype as the first major submission and milestone of work product, the author decided to focus on a risk that will allow him to hit as many major objectives as possible. With this in mind, the author has decided to develop around the classification of objects and the accuracy associated with this. This risk is listed as a medium risk to the project above.

The reason behind this was that it was deemed an achievable risk to mitigate against if it did go wrong. The submission assigned to this project by the customer (the university) would not leave enough time to fully develop a deliverable should something go wrong with risk 03 in the time outlined after necessary research in the planning section of this document. This is due to the technical complexity of the KeypointNet library and the experimental nature of the training process. Instead, the author has chosen to focus on the R-CNN system because he has a very limited prior knowledge of the library and believes he could mitigate against issues in an easier and more direct way.

Along with this reasoning, the R-CNN system is designed as the first section of the API. This reduces on the data the author may have needed to mock to build any other piece of the API. It also allows the beginning of development of an API and hosting this on a publicly accessible location. Furthermore, it would give him a perfect starting point for building the JSON structure that will be carried throughout the proposed system.

|  |  |
| --- | --- |
| Risk ID | Description |
| 01 | Not enough power to train models efficiently locally |
| 02 | Classification of objects with R-CNN not accurate enough |
| 04 | VM server could drop due to connection issues |

Figure 17 - Risks developed in prototype

In summary, the author has chosen to take a medium risk as the focus point of the prototype over a high-risk factor because it allowed me to mitigate and develop through several other risks listed. A high-risk factor is still being tested in the form of risk 01.

## 5.2 Design Artefacts

### 5.2.1 JSON structure

In planning for the prototype, the author first needed to outline a structure in which the system could pass data efficiently from one component to another. This comprised of what would be passed to the API through a post request, and what would be passed throughout and output from the API. These structures are listed below.

Figure 18 - Data to be passed to API

{

“image”:(base64 encoded image),

“objects”:(object list to detect),

“training flag”:(bool value)

}

Figure 19 - JSON that will be output from API

{

“image”:(base64 encoded image),

“RoI”:[(list of RoIs in image)],

“cropped images”:[

{

“cropped image”:(numpy array)

“saved filename”:(string),

“location”:(string),

“rotation”:[(float list)],

“object”:(string)

},

{…}

]

}

Each of these JSON structures may change slightly later in development however, for the initial prototype they contain all data the system will need.

### 5.2.2 Training images

In training the R-CNN model to recognise cups the author was required to hand annotate images of cups. This is a time-consuming process using an online tool called VIA image annotator. The annotations come back as a json file which can be used by the mask R-CNN training functions. Some of the training images and their annotations are listed below. In total, 15 images were annotated in this way, with a 12/3 split in images between training and validation.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Figure 20 - Training/Validation images and annotations

### 5.2.3 Prototype dataflow

The prototype has a relatively simple dataflow, however, it tests many different pieces of the infrastructure as well as testing different methods of training the required ML models. The dataflow is outlined in figure 21 below.

Figure 21 - Prototype dataflow

API (hosted on AWS EC2 VM)

Post Request

Base64 image convert

Detect RoIs

Crop Images

JSON returned

## 5.3 Issues in prototype development

As was to be expected at this stage of development, some issues were encountered when developing the initial prototype. The first major hurdle to this part of the project was the time it took from beginning of development to understanding the how to fully use the mask R-CNN library. This turned out to require a week’s work for the developer to develop a working knowledge of tensorflow, matplotlib, and the R-CNN library itself. This is something the author did not foresee during planning and unfortunately because of this, the project is now one week behind schedule.

Another issue that arose was the training of the model itself. The author had outlined this in the risk assessment and was a major risk he was testing with the prototype. The author acted out the mitigation plan that was in place and after noticing that the VM he had setup for the API was not able to use compute resources (more on this later), he transferred learning onto my local machine. Through this machine the author was able to pool more resources and was able to train the R-CNN model with the training dataset. A large issue with this was that while training, he was not able to complete work on his local machine. Depending on the training time this would not normally be an issue however, the model’s estimated training time per epoch was 2.5 hours. In completing 17 epochs, the local machine was unable to be worked on for 2 days.

This training time did not introduce unforeseen issues due to planning on the author’s part for this. It does however point to what will be a larger problem going forward where there is more pressure to have as much uptime on the local machine as possible. To mitigate this, the developer took 3 days to setup a separate EC2 instance that allows for better GPU compute resources. In addition, it increased the VM RAM from 1GB to 61GB, and 1 available core to 4. This dropped epoch train times to 12-25 minutes even when poorly optimized. More development time will be focused on optimizing training configurations to make full use of the new instance going forward.

The final hurdle to the prototype was the API instance in the cloud. When everything was complete and the code was ready to be moved into the VM, it was found that the VM did not have enough memory to complete an analysis on a single image. Due to the nature of EC2 instances, the author could have upgraded the memory at cost. Instead, he looked for a cheaper solution at this point in development to avoid large personal cost. He came across a technique called swap files that allow the allocation of additional memory at the cost of performance. This performance drop is caused by the offloading of memory to the hard disk. The performance drop is acceptable at this point in the development. A larger instance will be used for the final product demonstration.

# 6. Requirement control

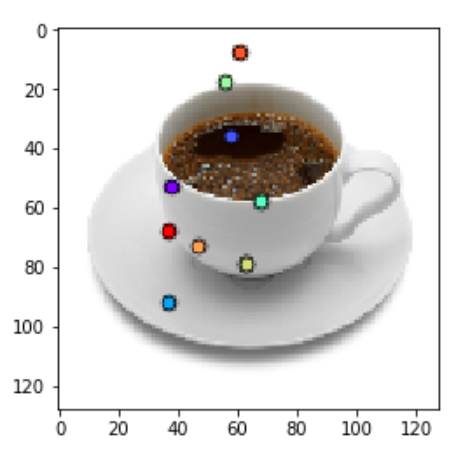
## 6.1 Final Requirements

|  |  |  |  |
| --- | --- | --- | --- |
| ID | FR/NFR | Requirement | Priority |
| 01 | FR | Application is standalone and does not require a browser | 1.234 |
| 02 | FR | Application takes single photograph as input | 1.968 |
| 03 | FR | User does not process the image locally | 0.260 |
| 04 | FR | Develop a public endpoint server that allows connection to ML models from anywhere | 0.409 |
| 05 | FR | System shall be trained using common objects | 0.451 |
| 06 | FR | System will hold standard 3D models of common objects | 0.536 |
| 09 | FR | The system will automatically communicate between different components of the pipeline | 0.352 |
| 10 | FR | The system will substitute 3D models in virtual space based on where the specified objects are in the input image | 0.256 |
| 11 | FR | The user should be able to show/hide metrics during execution | 0.579 |
| 12 | FR | The user should not have to leave the application during the process | 0.829 |
| 14 | FR | The application should allow the selection of objects to create in the 3D space | 0.680 |
| 15 | NFR | System will be developed using cutting edge open source frameworks for ML | 0.176 |
| 16 | NFR | System will have a way to report metrics including stage times, accuracy percentiles, etc. | 0.536 |
| 17 | NFR | System will not hold any data that isn’t marked as training data | 0.987 |
| 18 | NFR | System shall have a minimum of 70% test coverage | 0.581 |
| 19 | NFR | The API that receives and returns data should be secure | 0.736 |
| 20 | NFR | The application should be desktop only | 1.152 |

## 6.2 Requirements Evolution

Removed requirements:

|  |  |  |
| --- | --- | --- |
| ID | FR/NFR | Requirement |
| 08 | FR | System will recognise and report the depth of desired object relative to viewpoint |
| 07 | FR | System will recognise and report the relative rotation of desired object |
| 13 | FR | The application should have Virtual Reality support |



Unfortunately, due to the time required for the KeypointNet framework to be trained, these requirements were removed from the project. In total there was 4 weeks spent on developing the KeypointNet code for this project and testing this however the results from 5 days’ worth of training on the dataset were unsatisfactory. For example, in one of the test images, it was not able to accurately position the keypoints on the cup. Due to this taking up one month of development time, it had to be removed from the requirements however the code for the cup section is still in the client program and can be activated in the options menu. This will not change the test scenarios but will generate metrics based on the orientation estimate for the object in question.

In terms of the depth detection, I was not able to find an adequate replacement which would be able to detect the depth of an object in a monoscopic image and would be able to be developed and implemented in the time remaining.

The change in removing virtual reality support was simply because of the extra time taken up by the development of the keypoint network. This required a reevaluation during the beginning of cycle 3 and Virtual reality support was deemed by the stakeholders as being less important to the final product than the remaining requirements.

## 6.3 Project Plan Modifications

# 7. System Design

## 7.1 System Architecture

Figure 22 – Final high-level architecture overview

Input

Output

Unity client

Unity UI

C# Scene Builder

Virtual Scene render

3D Model storage system

API (AWS T2 Micro)

KeypointNet

Mask-RCNN

API entry

Collate JSON return

This was the architecture proposed in the planning document with only one major change. In order to retain the coupling of the API, the JSON container was removed from the final system. This change was made during development as the use of a separate class to write and read from rather than just passing the data in a pipeline was deemed a bad design by the stakeholders.

The system initially begins for the user in the Unity client and the user does not leave this client. The system was designed for ease of use for users and all processing of the data is done in the cloud. The API was built in python, while the unity client was all built using C#. Data was transferred throughout the system in JSON notation. In the Unity Client, all models and resources have been preloaded into the application allowing the user to do little-to-no configuration other than selecting a style of scene (for the purposes of this PoC) and submitting an image.

This system was designed and built for growth. With a modular structure, it is very easy currently for developers to add additional objects to be detected and implemented. This is assuming they have the trained models or compute resources in order to train. Within the API folder of the final application is all the resources required to train new objects, building on what has already been developed by the author. The API is currently hosted standalone on an AWS T2 Micro EC2 instance and intercepts calls using the python Flask library. In a production environment, a more scalable web server would need to be implemented as the T2 instance along with flask call only serve 1 request at a time. The author recommends increasing to an M4.large instance and implementing an nginx and uWSGI pipeline behind flask for simultaneous requests. As the models are all stored on the server, the web API does not require any other services to be running in parallel.

## 7.2 Interface Design

This section will outline the design considerations taken into account when developing the client application. The wireframes of the initial client are also below.

### 7.2.1 HCI Consideration

In consideration of the design of the interface, it was important that **H**uman **C**omputer **I**nteraction guidelines were taken into consideration. One of the defining guidelines was discussed by Molich and Nielsen in [27] wherein they outline 9 heuristics for categorising usability problems. These heuristics guide us towards problem areas and how we can avoid these by designing with them in mind. The 9 categories are outlined below:

|  |  |  |
| --- | --- | --- |
| **Problem area** | | **Description** |
| *1* | *Simple and Natural Dialogue* | Dialogs should only contain relevant and required information. The more irrelevant information the user interacts with, the less important relevant information becomes. |
| *2* | *Speak the user’s language* | The dialogue should be expressed in words and phrases that the user can recognise and understand. |
| *3* | *Minimise the user’s memory load* | The user should not have to remember information from one part of the system to another. Instructions for system use should be easily retrievable or visible when required by the user. |
| *4* | *Be Consistent* | Language used in one section of the application should mean the same in another. |
| *5* | *Provide feedback* | The system should always keep the user informed of what is happening within the application. |
| *6* | *Provide clearly marked exits* | The system should never confine users in situations that do not have any easily visible exit. |
| *7* | *Provide shortcuts* | Features such as tutorial dialogues or minimal entry fields that may be described as cumbersome to an experienced user should have ways in which these can be navigated easier for experienced users. |
| *8* | *Provide good error messages* | Good error messages are defensive, precise and constructive to the user experience. Error messages should never criticise the user and always blame problems on the system itself. |
| *9* | *Error prevention* | Careful design should reduce the errors the user sees. |

Figure - 9 Hueristics of HCI oulined by Molich and Nielsen[27]

As the API being developed has no interface for the purposes of this project, these 9 heuristics have not been taking into account for the design, with coding standards taking precedence.

### 7.2.2 Wireframes

Figure - Unity Client wireframes overview

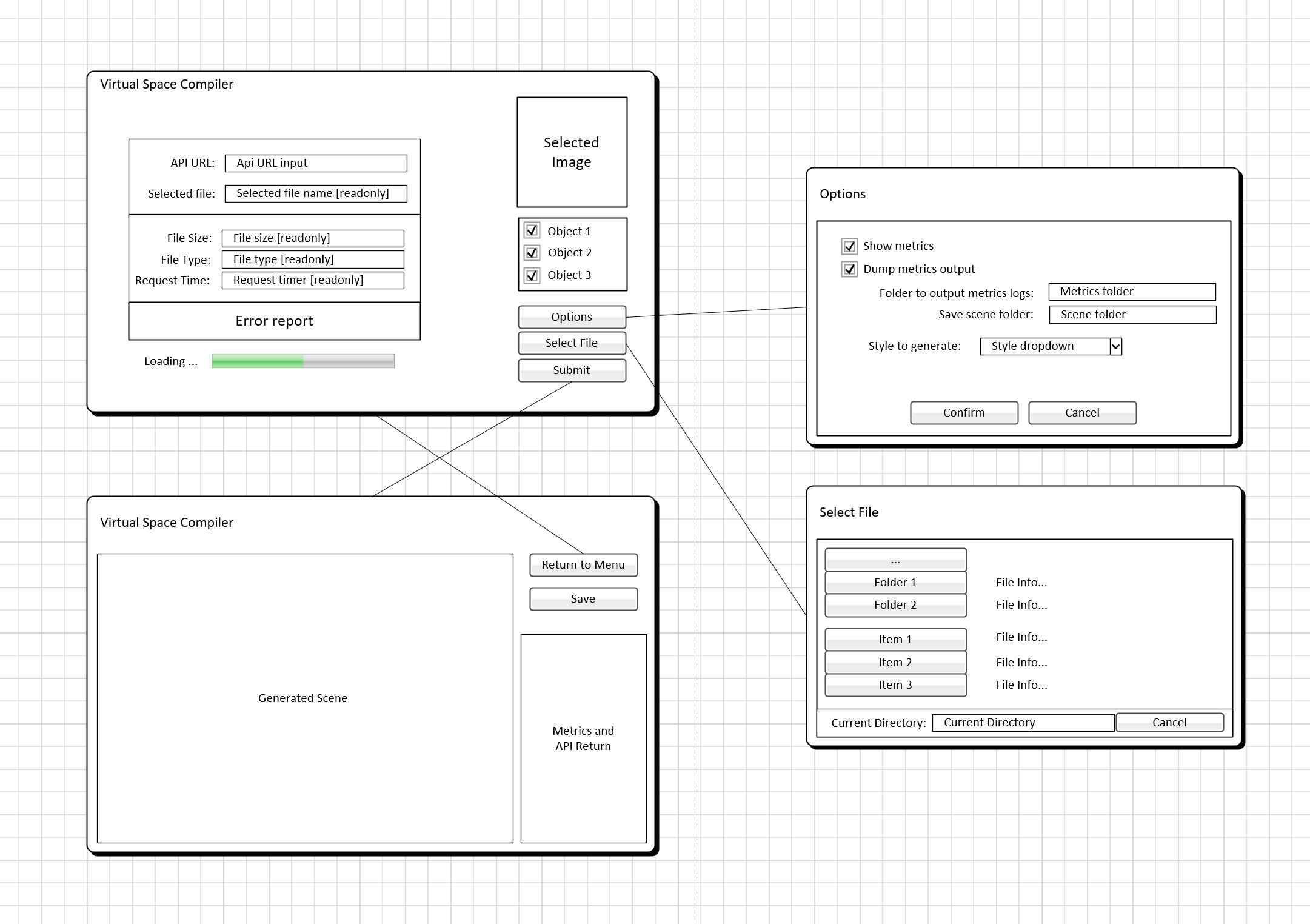


Figure 24 shows the wireframe interaction developed during project planning and design. These wireframes were generated in Microsoft Visio initially and added to during development for stakeholders to view the features that were being added to the program which were not originally stated but deemed to add value to the user. These wireframes were shown to and approved by the stakeholders outlined in section 2 of this document. These wireframes are guidelines used through UI development and do not fully represent the final design of the product. The application is comprised of two main application screens and two further dialogues.

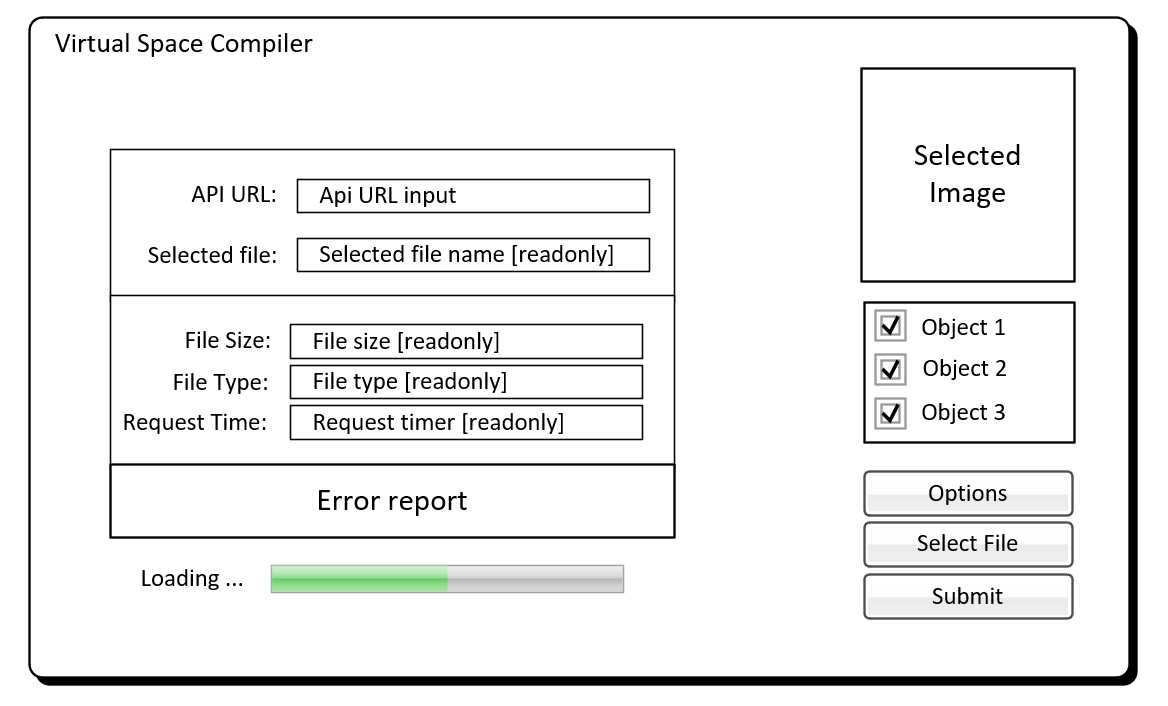


Figure - Main menu

The initial screen is the main menu of the system which quickly shows the information needed by the application in a clear and concise manner, in keeping with problem area 1 of figure 23. Here the user will input the URL of the API and select the objects they wish to detect. All other fields will automatically fill in keeping with the minimal user interaction goals set by the stakeholders. This menu will open 2 separate dialogue overlays: options and select file:

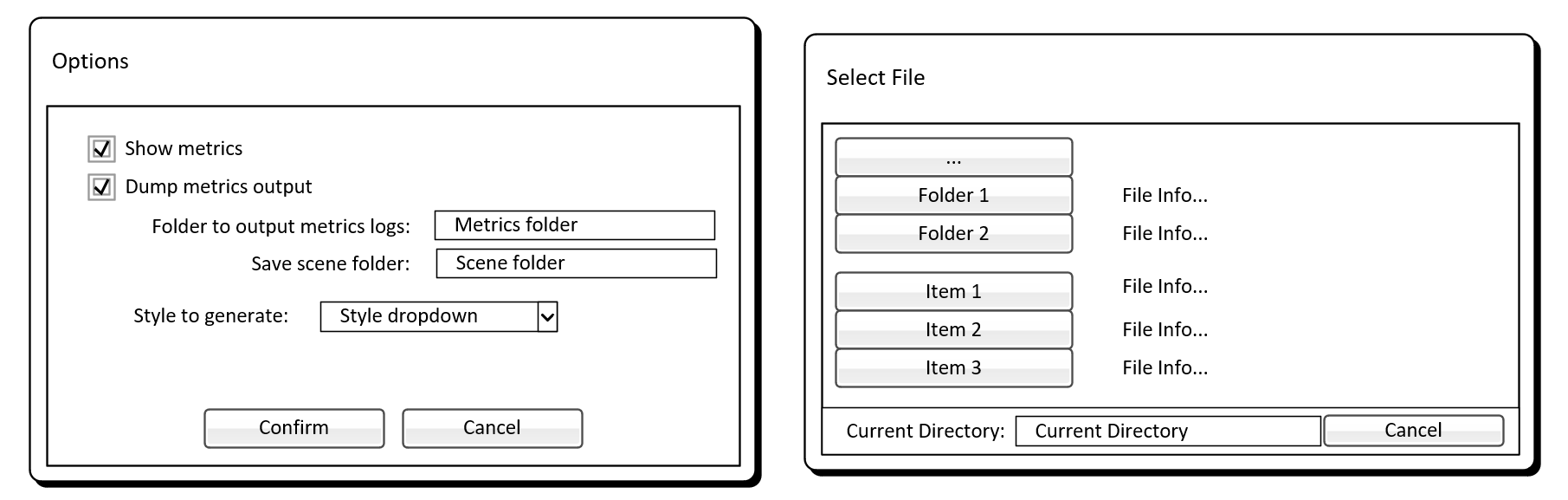


Figure - Options [left] and Select file [right] dialogues

The select file dialogue will open the user’s OS file select dialogue allowing the user to select which image stored locally to process. The options dialogue will allow the user to change style of scene, and to show/hide metrics. The metrics option will affect both main application screens, showing and hiding metrics panes as necessary. The error report section of the main menu will only be shown when an unexpected error occurs, adhering to problem area 5 and 8 of figure 23.

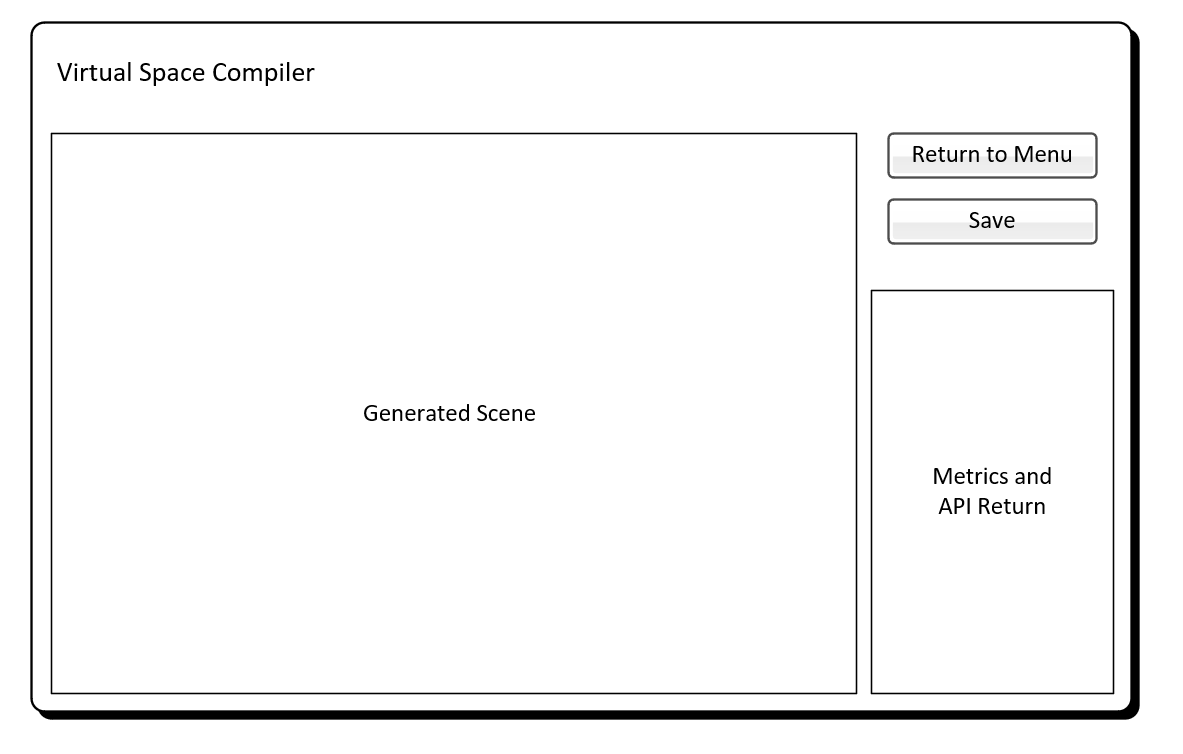


Figure - Generated scene

Figure 27 shows the final application screen. Here the user will be able to view the scene that has been generated by the API. The initial options for this phase of development are just to orbit the centre of the scene and zoom in and out. Further phases will look to allow the user to move and rotate items in the rendered scene dynamically to suit their purposes. The screen will also show the metrics from the system and the raw output from the API should the user select the show metrics option in the options dialogue (figure 26). The save button will output the scene object configuration in a way that will be readable for a unity developer. Due to the unity engine’s rendering method, the developers of this application are not able to save the whole scene directly without editing the underlying unity renderer code which is outside the scope of this project.

These wireframes have been developed in order to minimise the time it takes for the user to start the application and generate the virtual scene if they so desire. As such, the main menu only requires 3 inputs from the user: an image file, the URL to the API, and which objects to detect. In a production version of the application, the project stakeholders have indicated that the API URL will be automatically input further reducing user interaction. The reasoning for the API input is due to the use of a development server during this phase of development with no static address.

## 7.3 Data Support Design

Here the author will outline the considerations for the use and storage of data within the application. As one of the requirements of the project is for the application to not store any data of the user’s, the only data storage is metadata.

### 7.3.1 Consideration for security and validation

This application does not allow the writing of any data other than saving automatically generated log files to the user’s system. The API has been developed with only one endpoint which parses the request from the application. This request due to the use of ML models has built in soft failures from the tensorflow library. If any file other than a base64 image string is passed through to the endpoint along with registered objects defined by the developer of the API, a http error will occur. This reduces the chance of any attacks on the server.

As there is no SQL being used within the application due to all data being passed as JSON files, the server is not at risk of SQL injection attacks, the most common issues to hosted APIs. The main consideration for the security is the API itself which parses and validates all data that is passed from the application.

In terms of the validity of any data passed from the application, the data is all generated by the code itself other than the user image. In its current phase of development, we only look for .jpg images on the user’s file system to pass to the API as base64 strings. This is achieved using a custom-built file explorer in the desktop client. There remains one issue with this validation however it does not affect the data stored on the server (which is a total of 2 trained ML models) but affects the server itself. As we are exposing the API publicly currently and not hiding or fully validating the contents of the jpg on the client side, there is the possibility for a user to send a malicious file with the extension ‘.jpg’ and run their own code on the server itself. Due to the nature of this project as a proof of concept rather than a product or service, the fix for this issue was deemed out of scope to the stakeholders and has been placed in a further work section for future development phases.

### 7.3.2 current data design

For this project, the application configuration file have been stored in a document based database design. This allowed for decoupled configuration files for different sections which were added to through the development as needed. Each configuration file was formatted using the JSON standard. This kept each configuration easily readable at the beginning of the application without calling to a database each time the option for an object was needed. The reasoning behind this was also the small amount of estimated data that would be required for this phase of development. The files used to store configurations are outlined in a file name *ObjConfig.json* and structured as follows:

{

"object type": [

{

"objID":1,

"objName":"path/to/model",

"rel\_base":dist from pivot to base,

"style":"style"

},

{

"objID":2,

"objName":"path/to/model",

"rel\_base":dist from pivot to base,

"style":"style"

},

],

"opbect type": [

{

"objID":3,

"objName":"path/to/model",

"rel\_base":dist from pivot to base,

"style":"style"

}

]

}

Figure - ObjConfig.json structure

As you can see in figure 28, the configuration structure does not require enough data currently to require a full database design. This is due to all data being generated by our trained ML models.

### 7.3.3 Future data design

Before the system goes into production, a more efficient way for storing and changing configurations remotely would need to be developed. This would be required by larger businesses to continue use of the product as they would need to customise styles and secure the users of their system. Each system database would be deployed for the business ensuring no cross-contamination of business data and reducing the effect of a data breach should it happen. As such, a high-level configuration database ER diagram has been outlined below in figure 29.

This database configuration would allow each business to deploy and manage their own database and API server (possibly a future revenue stream for the development team to deploy and manage these servers). This also would restrict API access by making use of generated API keys, reducing the chance of server exploitation and increasing access logging.

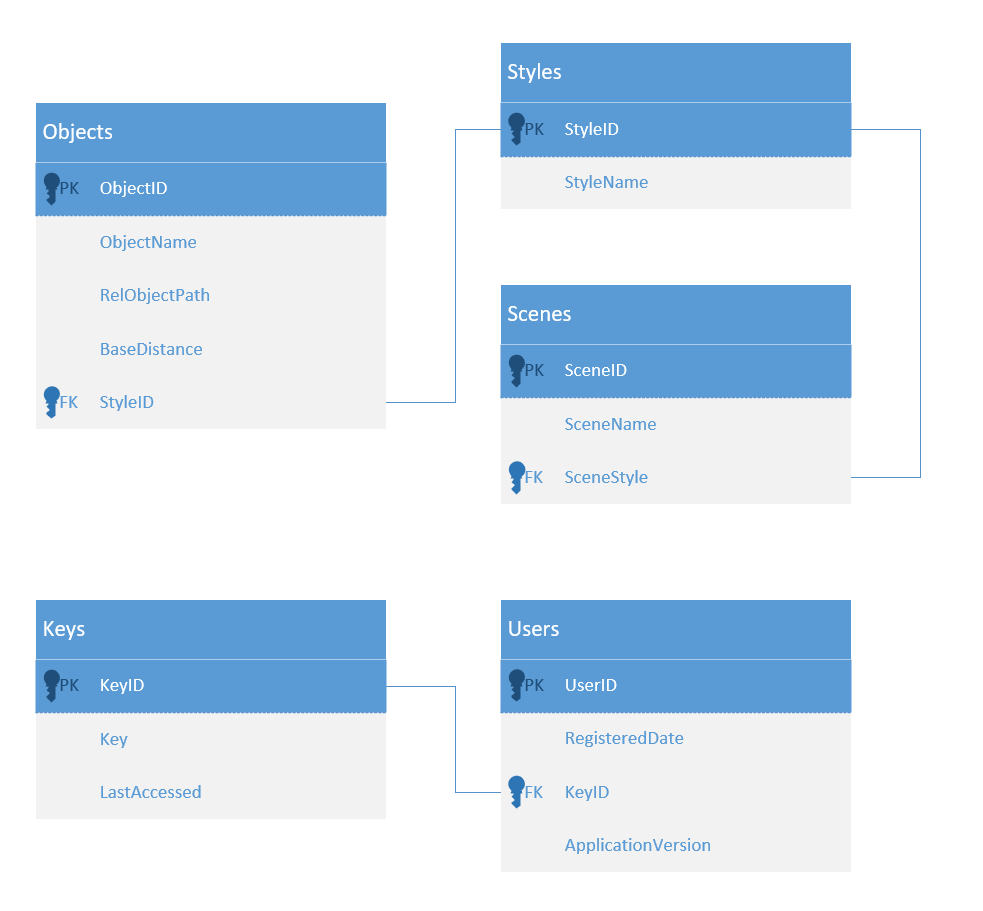


Figure - Future high-level ER diagram

## 7.4 User Interaction

### 7.4.1 Data flow diagram

Figure - Data flow diagram

Illustrated above in figure 30 is an overview of data flows in the application and data stores. As you can see, no major user data is required as outlined in section 7.3. There is also no external entities required to complete the system – it is a fully self-contained system in it’s current form. The application focuses more on the passing of data rather than the storing.

### 7.4.2 Use case model

A use case model was developed by the author at the beginning of cycle 2 of development. This was a way for the developer to visualise the tasks the system would be required to carry out to complete all requirements.

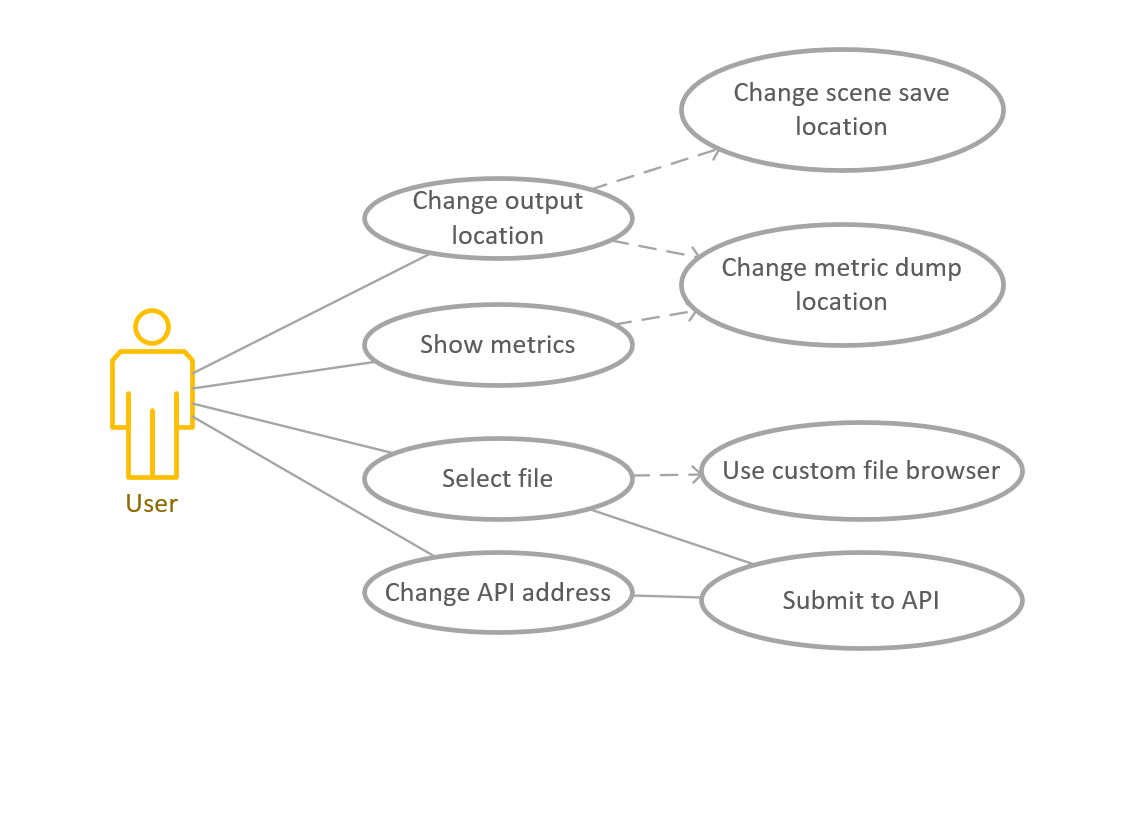


Figure - Use case model

This use case model was checked at each cycle end to ensure the developer was approaching the user options required by the stakeholders. As you can see in figure 31, there is minimal user interaction with the system. This is to emphasise the automated systems the project intends to prove.

## 7.5 Additional

### 7.5.1 API Information process

As the API is doing a lot of data changing and processing, the author in this section outlines the processes in which the data will be parsed and estimated. The author here will refer to activities outlined in figure 30.

The desktop client application in its current form will output a network request body of two pieces of information:

* A base64 string encoded image
* An array of strings detailing which objects need to be detected

The second of these is a list of strings currently to allow better readability in early stages of development and in the attached code document. This will – in future product versions – become simply an array of integers representing IDs of objects. These two pieces of information are packaged into a JSON request and sent out as a HTTP PUT request to the API (This was because of a bug in Unity’s network interface).

Once this request is received by the API it is parsed into it’s two objects on the server side. From here the base64 image is decoded from base64 to a numpy array [28]. This array is essentially a list of each pixel and its value in integer format which is readable by our ML algorithms. In parallel to this, we look at the request to see which objects the client want’s the API to detect. This is just taken from the second of our request parameters.

At this point, the image is passed into our objects trained RCNN models (one per object for more accurate classification). The training of these models comes from separate files and we only need to store a single model rather than the whole training dataset etc. This allows everything to be held on the server without large storage systems. The RCNN model sorts through each image and decides which pixels in the image – or pixel groups – belong to the specified object. More on this process is found in section 3 of this report. The model returns a region of interest field for each object.

This region of interest is then cropped from the image for separation to the keypoint network for orientation detection. Along with cropping the image, we gather the location of the image relative to the camera position. This returns 4 parameters:

* True x: the actual x coordinate of the centre of the object
* True y: the actual y coordinate of the centre of the object
* Relative x: the relative x coordinate ie. how far across the image in percent is the object centre
* Relative y: the relative y coordinate ie. how far across the image in percent is the object centre

These data points are used by our client to properly position the object in the camera view.

The cropped image of the specified object is then passed to the KeypointNet subsystem as shown in figure 30. This keypointnet requires our cropped image to be converted into a tensorflow proprietary format called a .tfrecord. This format is a way for tensorflow to more efficiently process large amounts of data. We pass through the cropped image and the trained KeypointNet model detects the keypoints of the object it had trained (more on this process is found in section 3), and returns 3 datapoints for each:

* Average depth of object
* Object orientation
* Keypoint positions

These three datapoints are packaged into a JSON return string along with the RCNN returns for each detected object and are sent back to the desktop client for scene placement. This is all the processing required by the system, offloading the processing of the image and models to the cloud, and reducing strain on the user’s system. This means that the system can run on systems that would otherwise not be able to use ML model processing. It also allows for scaling of the system speeds as we can simply increase the available CPUs in the cloud to reduce the time for the request to be processed.

# 8. System Implementation

## 8.1 Reflection

In reflecting back now at the end of this project, the author has returned to the implementation plan outlined in section 3 of this report. The plan itself was originally very stable and – although there was a very large project scope – was entirely within the grounds of reason. One thing that has been noticed when looking back on this plan is that not a lot of time was give for external work, and submissions required for other modules. This became too overwhelming starting in March of 2019 where much of the developer’s time was taken by other module submissions. This pushed the project timescale back as outlined at the end of section 6. This was an unfortunate mistake in the planning process, and a learning experience for the author.

Another major issue with the plan was the training of ML models. It was estimated that these models would take time to train (which was accounted for in the original plan structure) however these estimates were incredibly optimistic looking back. One model – which as previously mentioned was not in the final product – took 5 full days of training time which blocked the development. Looking back, the spiral development methodology was perfect for this project as it did involve a lot of risk and allowed the developer to focus on high risks each cycle instead of putting them off. This meant these risks were easily mitigated against.

## 8.2 Frameworks and tools

This project made use of heavily experimental frameworks and as such, the author kept many other tools to what they were comfortable developing and supporting. This section aims to break these down and explain the rationale behind each

### 8.2.1 Tools

Task control was completed throughout the project. This was done through the use of GitKracken, supplied through github student. This task tracking was to ensure the initial goals outlined by the stakeholder were achieved and tasks were completed on time. Below is a screenshot of this board.

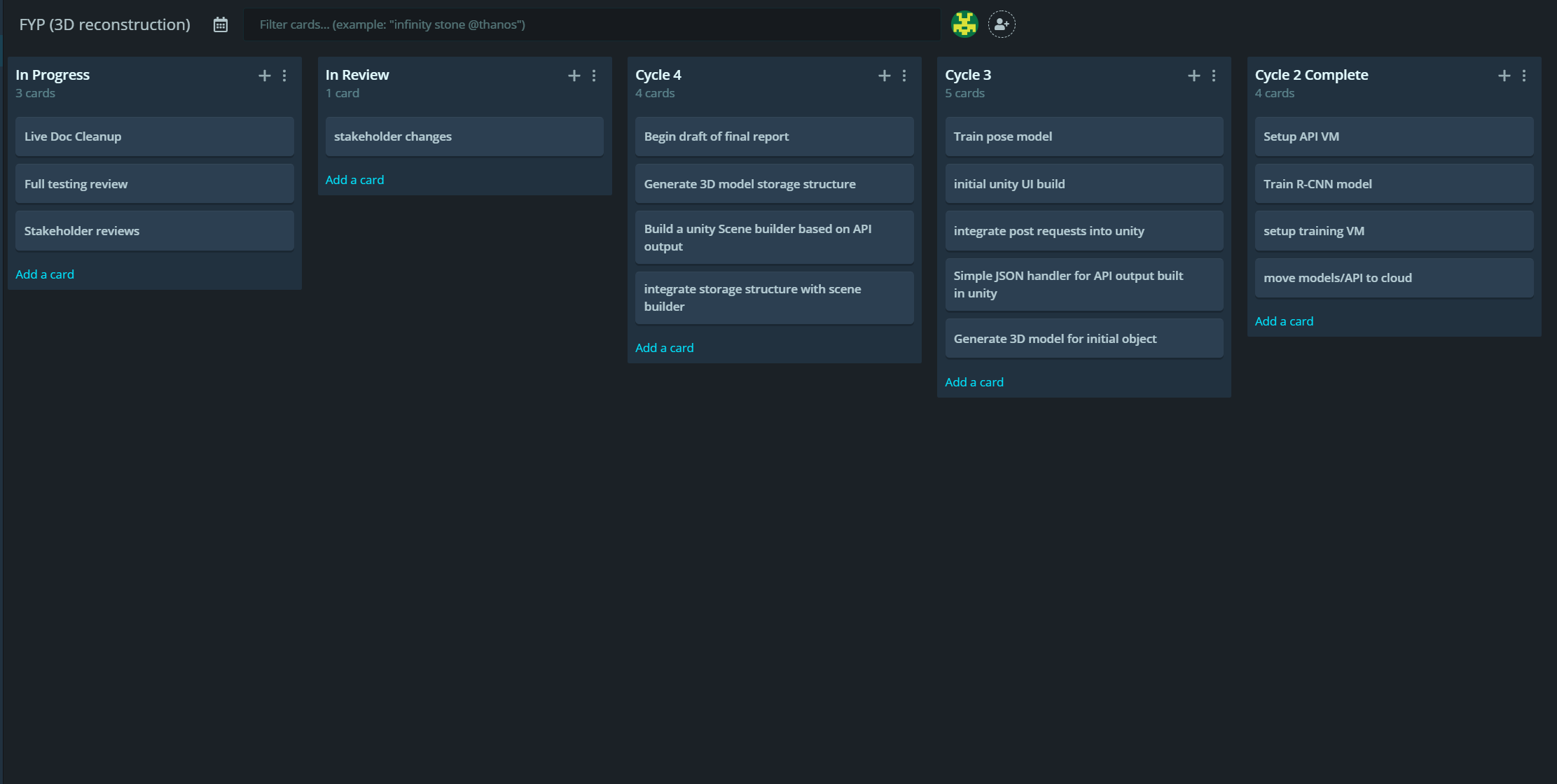


Figure 32 - GitKracken task board

The board shown in figure 32 is broken into a task list, in progress, in review, then the tasks are assigned to the cycle they were completed in. This board was public and could be seen by stakeholders to track progress.

For the main development of the data processing, building of the datasets for training, and image detection validation, the developer use Jupyter Notebooks. This is a free lightweight program which allows modular development of python programs and is widely used in ML development. Much of the backend development of the solution depended on this modularity to increase efficiency of development.

For the frontend of the project, the unity physics engine was used. This was the authors main development environment through industrial placement and as such, they were very comfortable in the building and optimizing of desktop applications through this program. Unity works in tandem with Microsoft Visual Studio for the interaction code for the program. With these 3 main tools (Unity, MS VS, Jupyter Notebooks), the whole application pipeline was developed. The deployment of the application to the cloud was done completely manually by the author as Amazon AWS made this simple to do without any extra technical training in a new system.

### 8.2.2 Languages

The languages used in this project were all well known by the developer prior to beginning the development. The python programming language was used for backend development for its flexibility and scalability for API development. This was used in conjunction with the flask library to create a web API that allows open access to the ML models when given the correct data.

For the frontend code, C# was used. The rationale behind this was that it was the language the developer was most confident in, having over 6 years of development experience in this language. C# is also the default backend scripting language utilised by the unity engine.

### 8.2.3 Frameworks

This project was overall a test of building a pipeline to use two experimental frameworks which have been discussed throughout this report. The author chose both frameworks as they initially showed a way to achieve monoscopic image recognition and detection when combined. The first of these frameworks is the Mask RCNN framework and algorithm developed by Matterport for object recognition and isolation. This framework was required to detect and isolate each object of the test image from the busy background. This was then used to place the objects in the virtual scene relative to the camera.

The second framework that was key was the KeypointNet framework. This was built upon the tensorflow libraries to automatically detect key locations on each image and relate these to the image in order to respond back with orientations and depth. In the end, this network took too long to train to be viable for current use. This is where the solution fell down in the end, this is expanded upon in section 11.

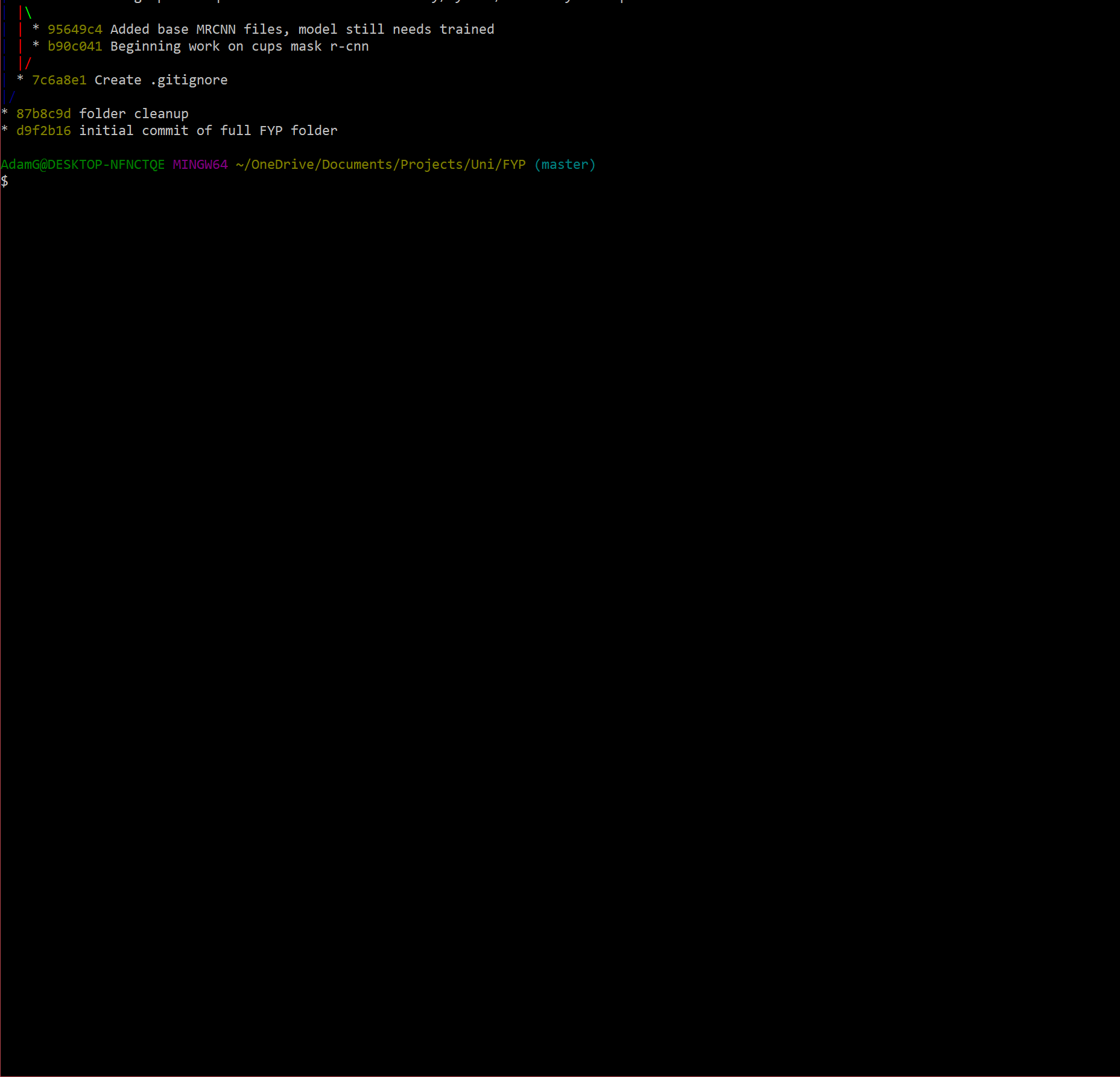
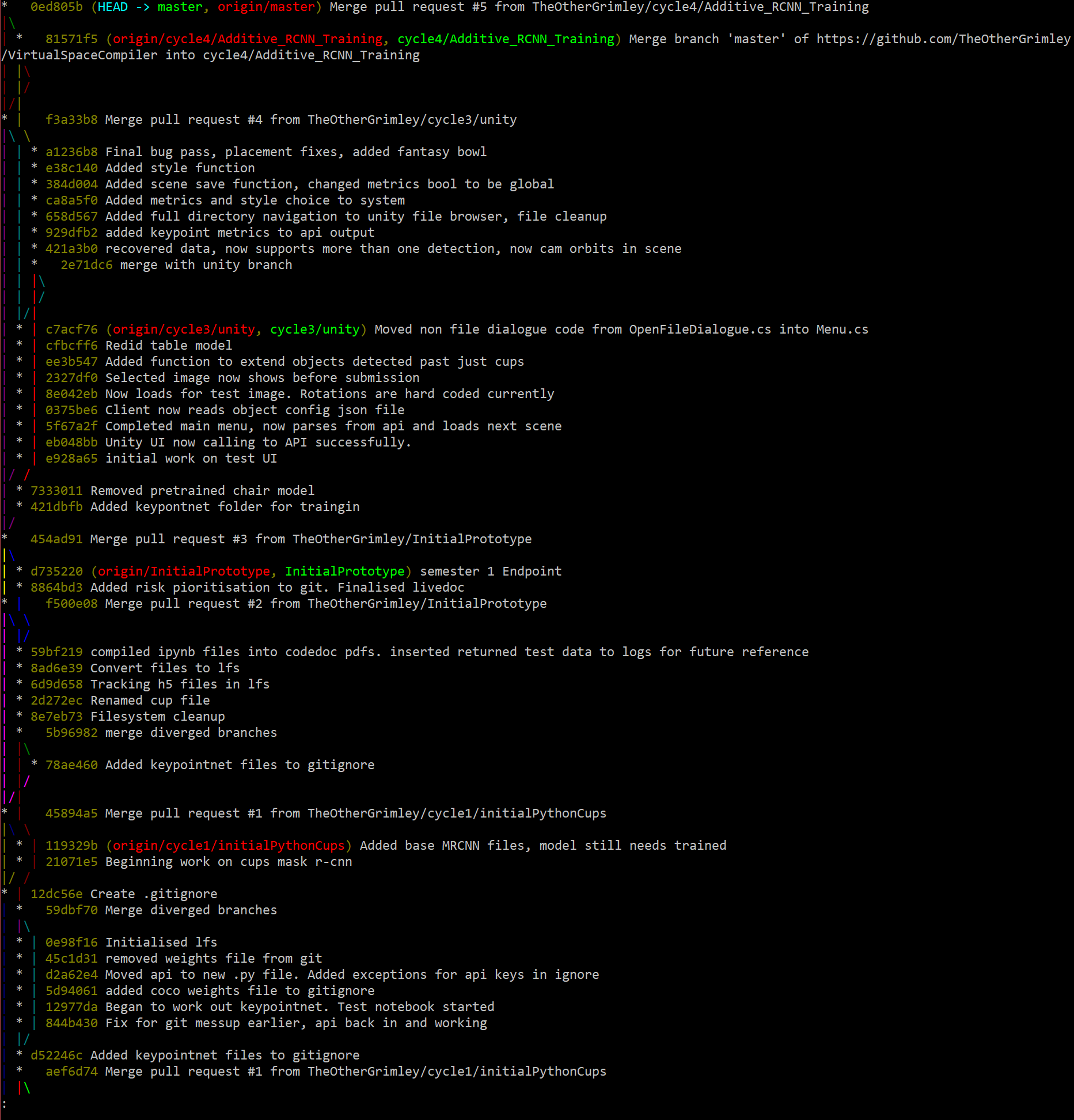
This project also made use of the flask framework in order to build out a REST API for backend storage and data processing. This was due to flask’s ease of use and use in production environments relevant to the author. In using flask, the author not only gained vital knowledge in the use of relevant web hosting technology but also learned a lot about building network technologies. The project also passed all data as JSON, a standard which was built for javascript but has become one of the most versatile standards for sending and parsing data. Due to this, the LitJSON package was used within unity which is a group of JSON handlers that bolsters the unity engines built in methods for working with JSON data.

## 8.3 Source control

In terms of source control, this project made use of git. Other source control technologies were considered such as mercurial or subversion however git was shown to be the easiest to implement as well as the most well-documented standard in the authors opinion. Due to the developers comfort in terminal, the bash version of git was used throughout the project.

A proper git flow branching strategy was put in place by the author to help in the project planning and implementation. The project was broken up into it’s separate cycles and merged into master at the end of each cycle. This flow did admittedly break at the end of cycle 3 where it was deemed by the developer that it would be much easier to tie into cycle 4 than go back to master as the code required to finish cycle 3 was at the beginning of cycle 4. The project was finalised with a total of 51 commits to a private github repo.

Figure - Git commit log



## 8.4 Code volume

This project was split into 2 main products (API and Client). Each of these have been broken down below

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Classes** | **Methods** | **Lines** |
| *Api.py* | 2 | 14 | 232 |
| *Cups.py* | 3 | 5 | 174 |
| *Bowls.py* | 3 | 5 | 174 |
|  | **8** | **24** | **580** |

Figure 34 - Backend code analysis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Structs** | **Methods** | **Lines** | **Variables** |
| *FileDialogue* | 1 | 11 | 156 | 15 |
| *ImageLoader* | 0 | 2 | 33 | 1 |
| *Menu* | 1 | 7 | 116 | 5 |
| *Metrics* | 1 | 4 | 102 | 8 |
| *Options* | 0 | 4 | 38 | 3 |
| *Orbit* | 0 | 1 | 48 | 2 |
| *OrientIn (not used)* | 0 | 1 | 20 | 4 |
| *Rotate* | 0 | 1 | 15 | 3 |
| *SceneBuilder* | 3 | 8 | 217 | 9 |
| *SceneData* | 4 | 3 | 71 | 3 |
| *ToggleInit* | 0 | 1 | 17 | 0 |
|  | **10** | **43** | **833** | **53** |

Figure 35 - Client code analysis

On a code metrics generation on the frontend, Visual Studio’s tools returned figure 36.

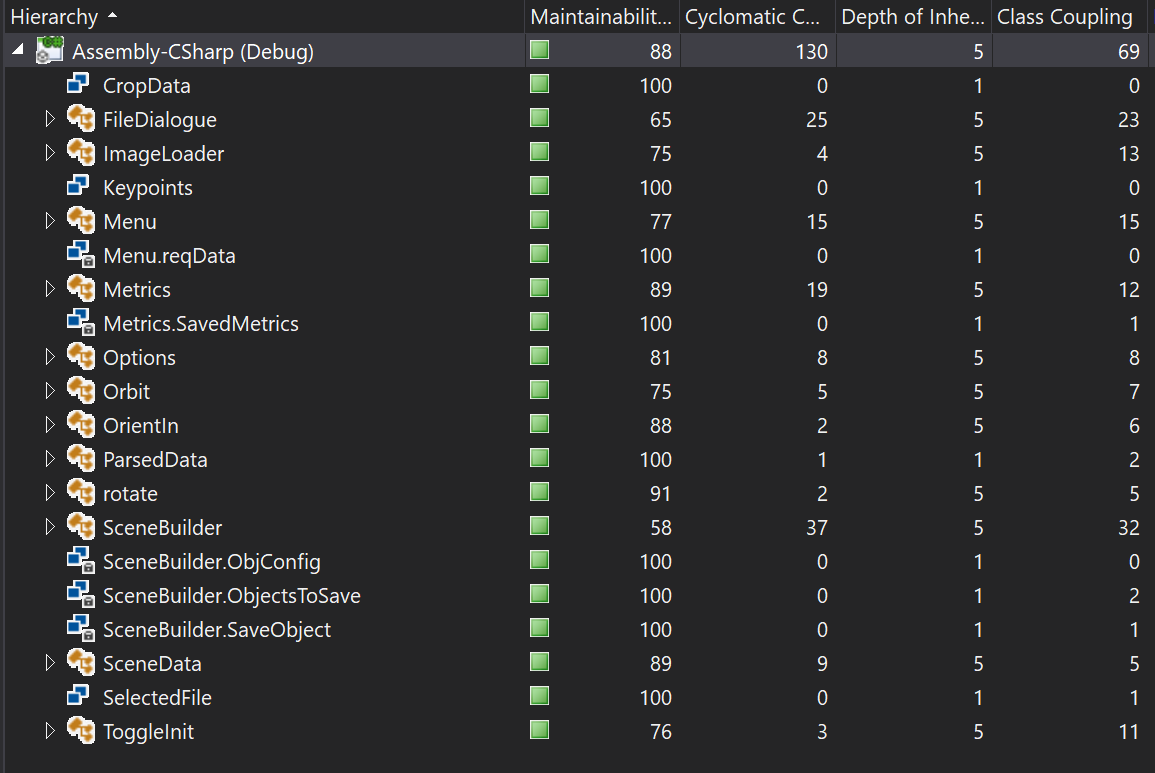
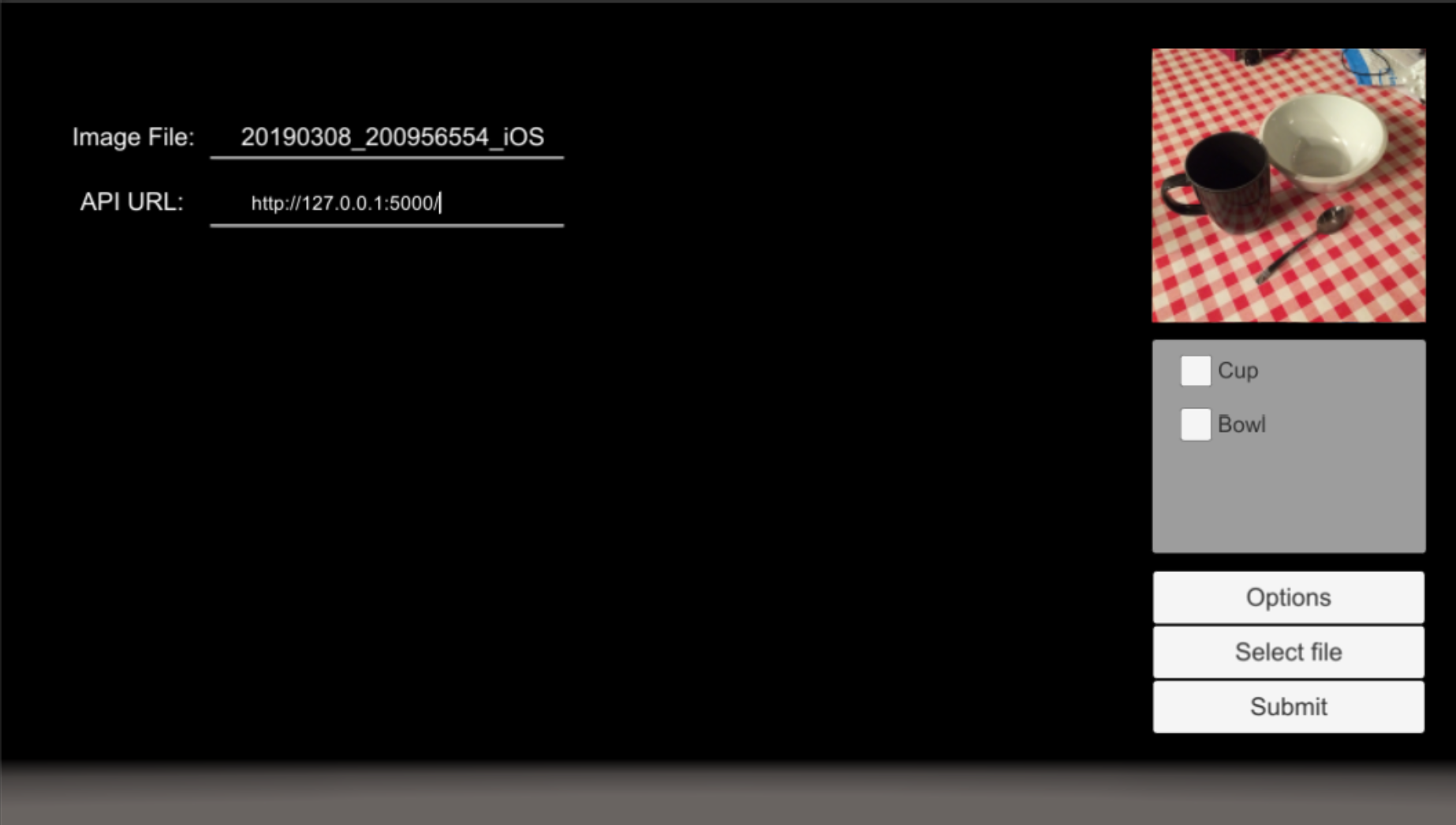
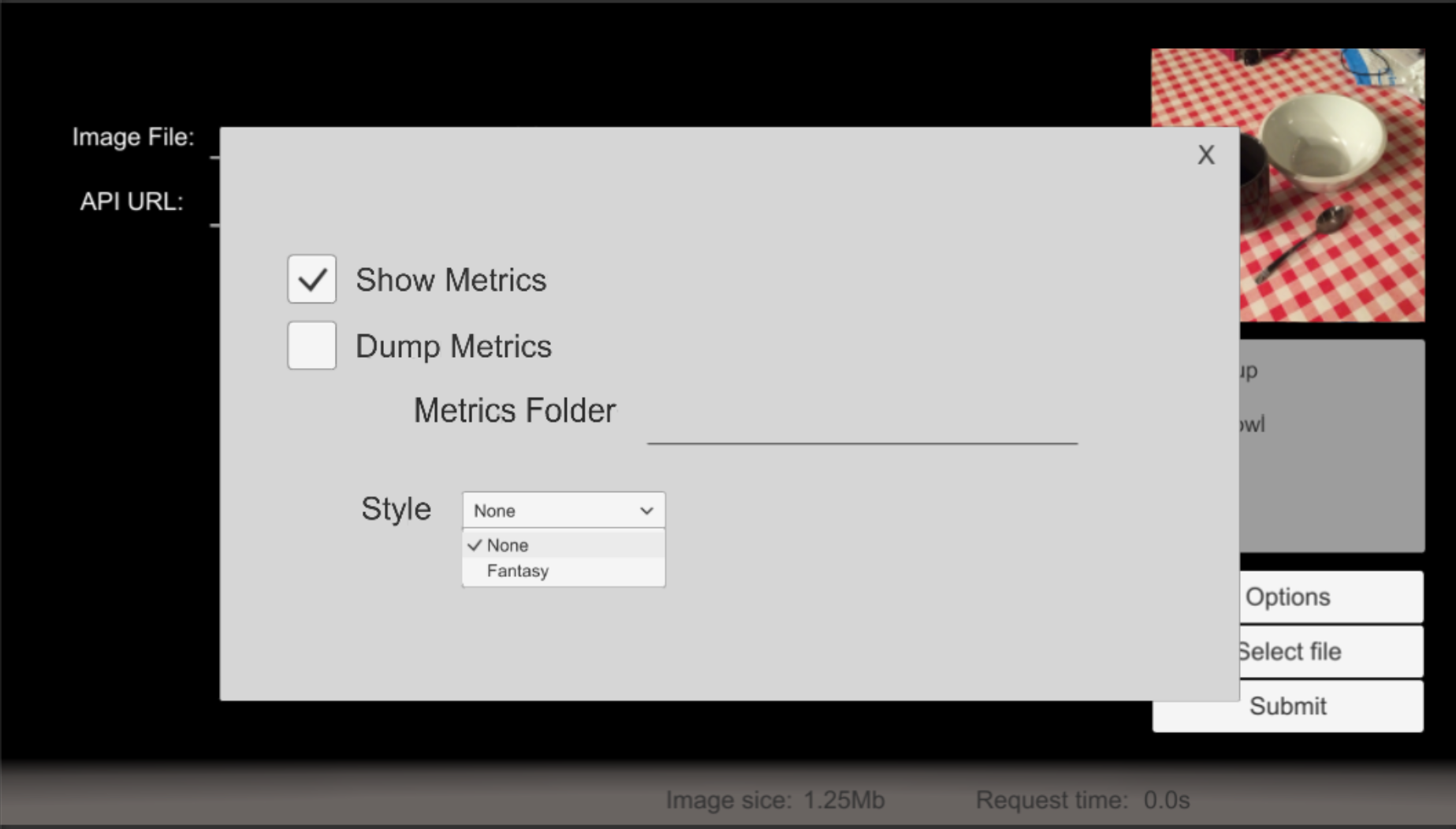


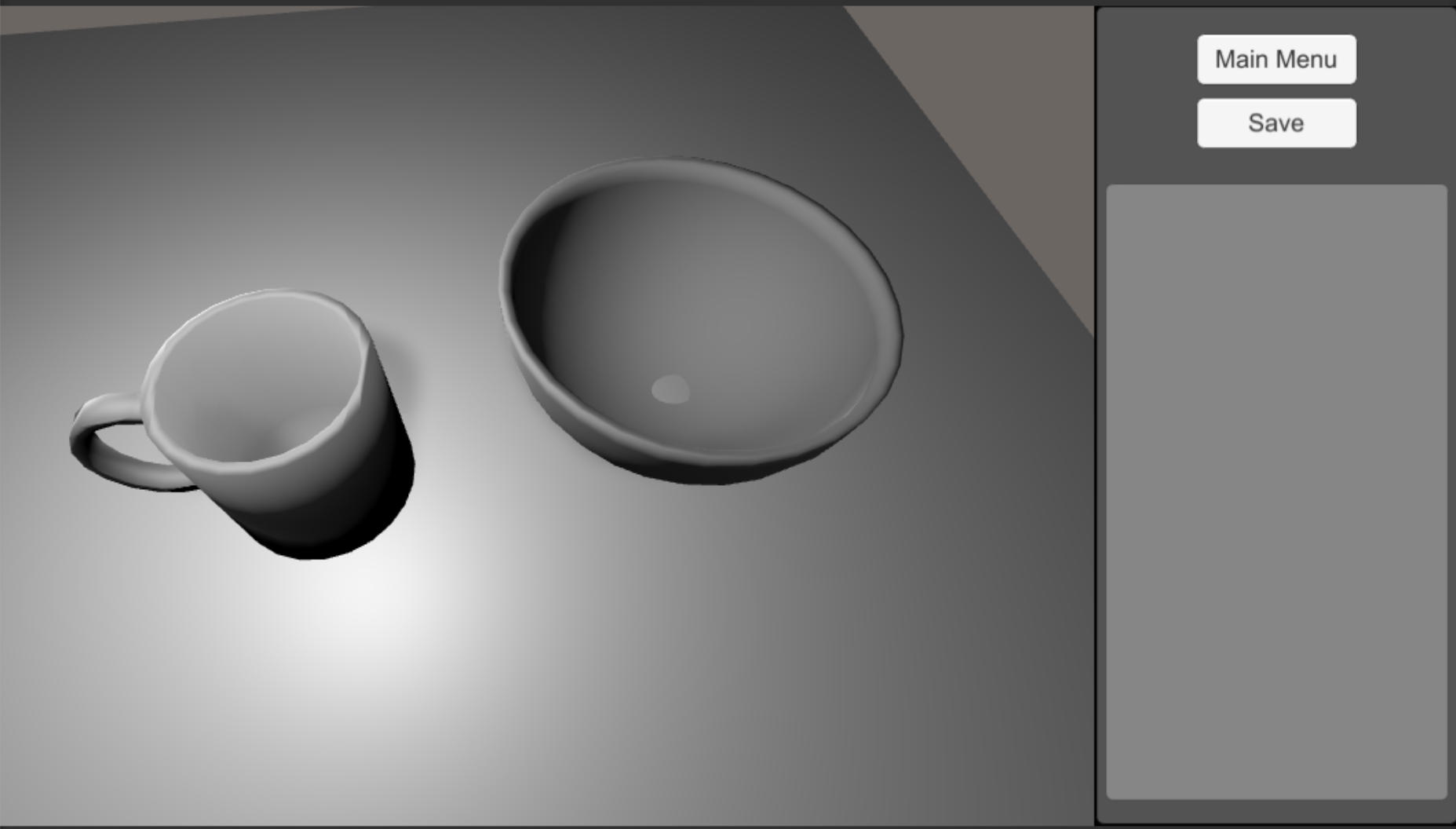
Figure - Code metrics

## 8.5 System Walkthrough









## 8.6 Security

# 9. System Verification

## 9.1 Reflection

## 9.2 Verification Plan Results

## 9.3 Confirmation of Requirements

# 10. System Validation

## 10.1 Reflection

## 10.2 Validation Results

## 10.3 Future Work

# 11. Conclusion