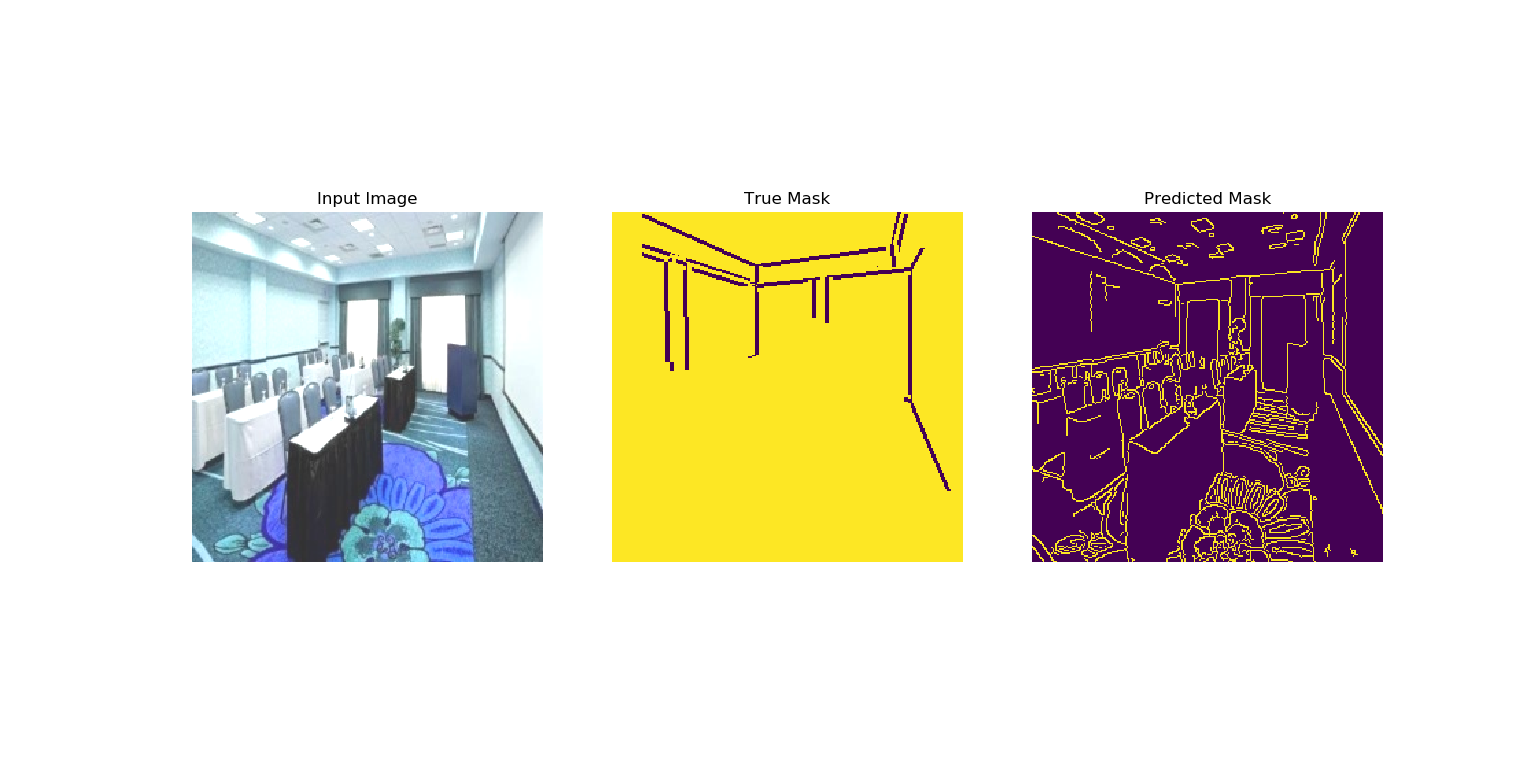
**1. Abstract**

**2. Introduction**

**1.1 Background and Motivation**

This project aims to reconstruct the 3D model of room architecture from photographic images. This has many applications, such as measuring room dimensions and interior design. The greatest challenge in this process is extracting the relevant edges in photographs. Due to the wide variety of noise present in photographs, a deterministic approach to the problem, such as the Canny edge finder, proved to be inadequate based on past projects (Figure 1). This is because classic algorithms fail to distinguish between different edge types.



*Figure 1 – Results from Canny Edge Detector (OpenCV)*

**1.2 Objectives and Scope**

As such, machine learning is proposed as a possible alternative for detecting structural edges in photographic images. This project therefore aims to establish whether machine learning is viable, and whether it has a higher accuracy than traditional algorithms.

**1.3 Layout of Report**

**3. Literature Review**

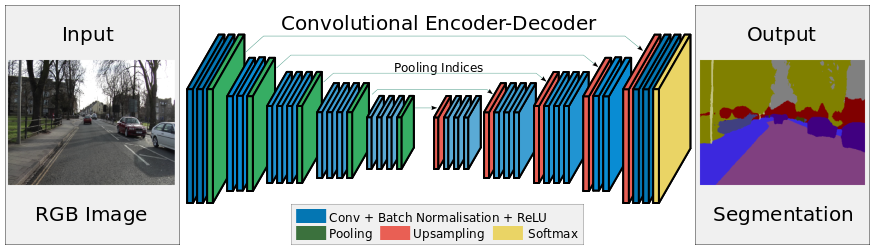
**3.1 Overview**

Recent developments in machine learning have mainly revolved around deep learning and it represents an opportunity for machines to perform tasks that have traditionally been impossible (Marr, 2018). In particular, Convolutional Neural Networks (CNN) and Fully Convolutional Networks (FCN) has progressed significantly, leading to huge improvements in image processing capabilities. FCNs are a natural progression from CNN and they differ from CNNs in their last layer. While CNNs generally have a linear output layer for classifying an entire image, FCNs have a convolutional output layer that attempts to classify each pixel in an image (Jonathan Long, 2014). This has caused FCNs to rapidly gain popularity as it can potentially identify multiple objects in a single image as well as the location, size and shape of each object. This process is known as image segmentation and has applications in many fields, including smart cars.

**3.2 Fully Convolutional Networks**

There are a few notable FCN architectures that are readily available. The more successful ones include SegNet, U-Net, DenseNet, and many more (Shah, 2017). More details regarding some of the architecture is provided below.

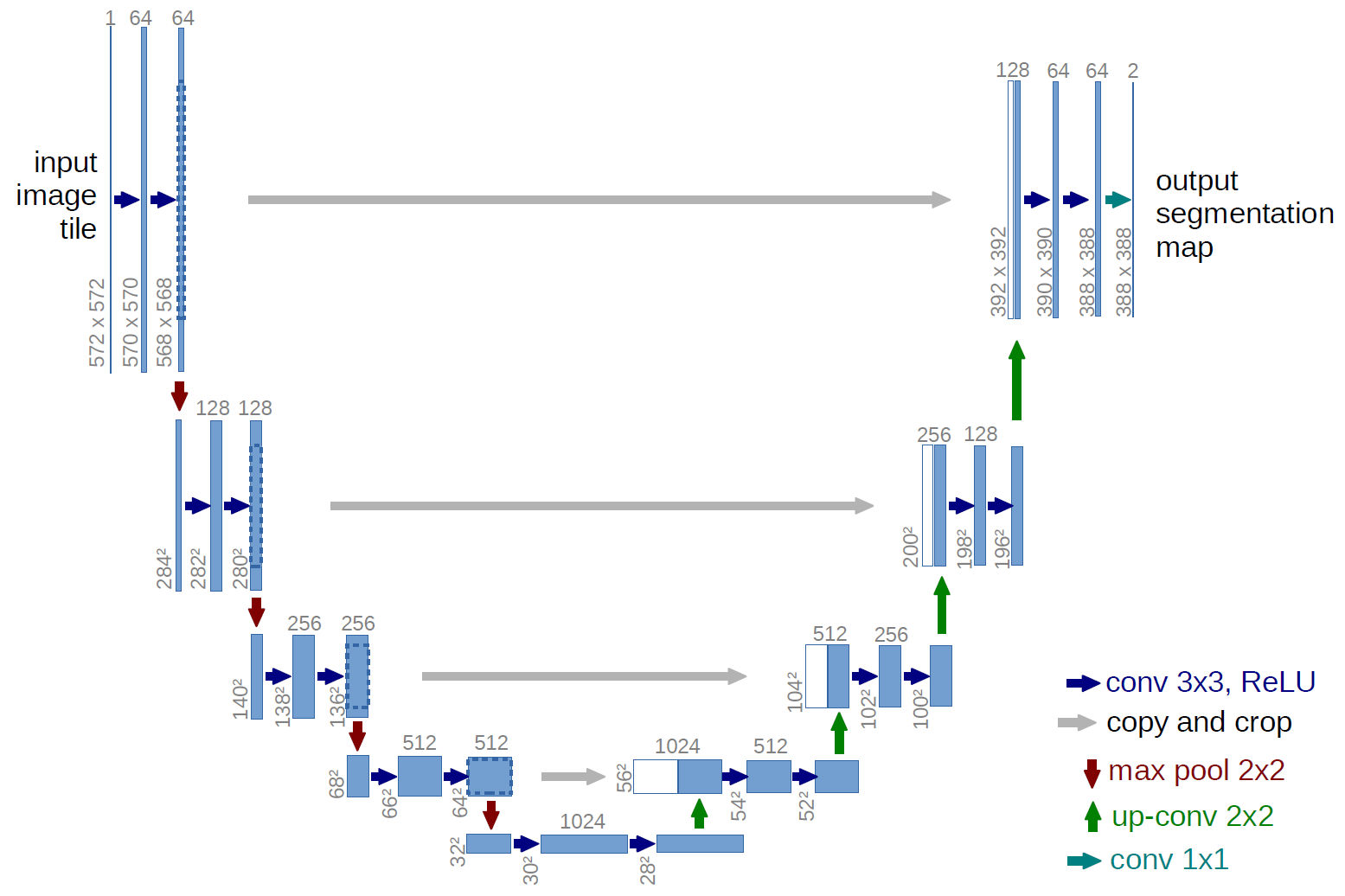
1. SegNet architecture (Vijay Badrinarayanan, 2015)



*Figure 2 – SegNet architecture*

SegNet uses cascading modules of encoders followed by cascading modules of decoders to extract key features from an image. Encoding modules consist Convolution, Batch Normalization, ReLU (activation function) and Max Pool, while decoding modules consist Max Unpool, Convolution, Batch Normalization and ReLU.

1. U-Net architecture (Olaf Ronneberger, 2015)

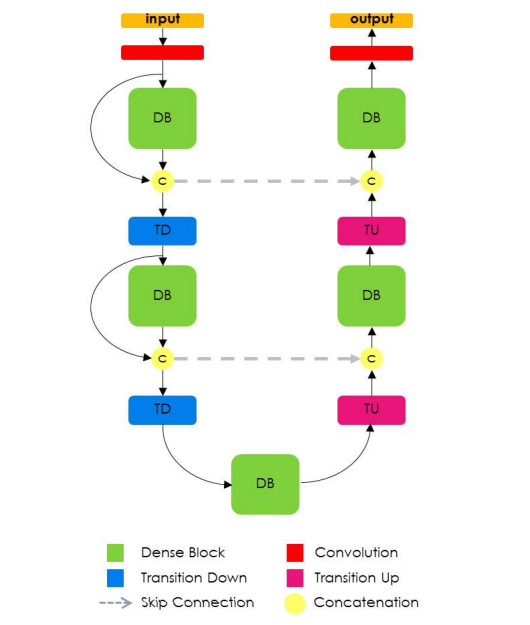


*Figure 3 – U-Net architecture*

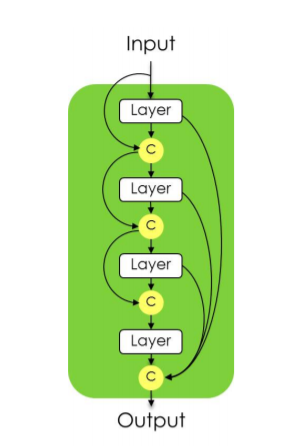
U-Net is similar to SegNet in many ways except 2 key differences. First, it uses Transposed Convolution instead of Max Unpool for upsampling. This creates additional parameters for the model to train on. Second, the output from each encoder block is cloned and concatenated onto the input of each respective decoder block. This allows some features that might otherwise have been lost to be captured by the decoder blocks.

1. DenseNet architecture (Simon Jégou, 2016)

DenseNet is similar to U-net in almost every aspect except for its building block, the Dense Block (Figure 5). The Dense Block is made by concatenating the input with the output of each layer and concatenating the outputs of all 4 layers together at the end.



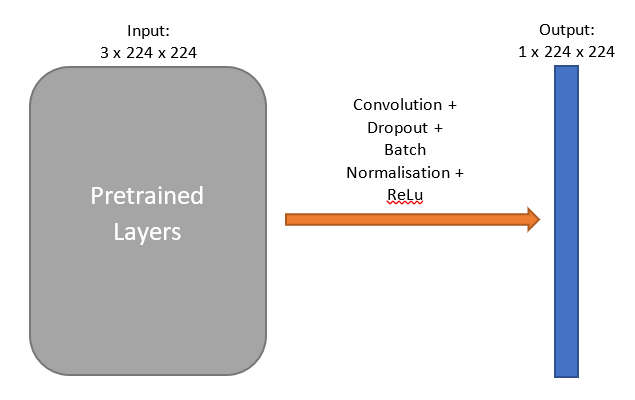
*Figure 4 – DenseNet architecture*



*Figure 5 – A Dense Block in the DenseNet architecture*

1. Transfer Learning using pre-trained FCN-RESNET101

FCN-RESNET101 is a Fully Convolutional Network model with a ResNet-101 backbone. The model has been pretrained on a subset of the Common Objects in Context (COCO) 2017 dataset. The pretrained weights were used, and only the last few layers of the model were modified and retrained for the purposes of this project.



*Figure 6 – Transfer Learning model using FCN-RESNET101* *architecture*

Testing is required to determine which architecture would work best for the purposes of this project since it was difficult to determine at first glance. For practical purposes, only SegNet, U-Net, DenseNet and FCN-RESNET101 were chosen as candidates. In order to implement these architectures, a suitable language and library must first be chosen.

**3.3 Machine Learning Libraries**

Most machine learning programs are written in Python due to its simplicity and the tremendous amount of open source libraries and resources supporting the language. In recent years, 2 machine learning libraries have become widely popular:

1. Tensorflow – developed by Google
2. PyTorch – developed by Facebook

While both libraries have its strengths and weaknesses, PyTorch was chosen the purposes of this project as it was a lot more memory-efficient in implementing the neural network, a condition that was necessary due to hardware constraints. In addition, it was more intuitive than Tensorflow when it comes to implementing non-sequential neural networks, such as U-Net and DenseNet.

**3.4 Image Processing Libraries**

Another tool that is widely used in image processing is Open Source Computer Vision Library (OpenCV).

**4. Methodology**

**4.1 Overview**

The machine learning model chosen for this project is the Convolutional Neural Network (CNN). CNNs are widely used for image processing due to its ability to extract features from images. The input to the model is a 256 x 256 RGB image and the expected output is a 256 x 256 grayscale image containing the structural edges found in the image.

The process of training the model is divided into three parts:

1. Labelling of data
2. Data Augmentation
3. Data Processing
4. Training the Model

The images that were used were sourced from a scene-centric database used in the LSUN Scene Classification Challenge. Approximately 5,000 images from categories “bedroom”, “classroom”, “conference room” and “living room” were selected as the dataset for this project (Figure 7). The rationale for choosing images across multiple categories was to increase the variety of objects in and the layout of the room depicted in each image. This reduces the likelihood of overfitting and improves the model’s ability to recognise rooms that are not covered in the dataset, such as storerooms.

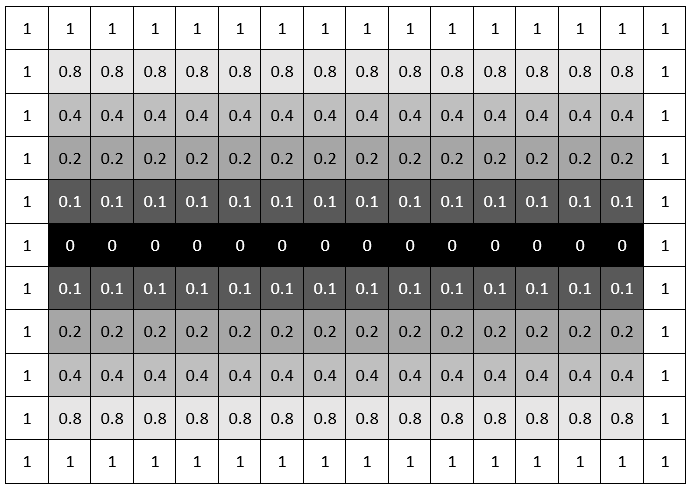
**4.2 Data Labelling**



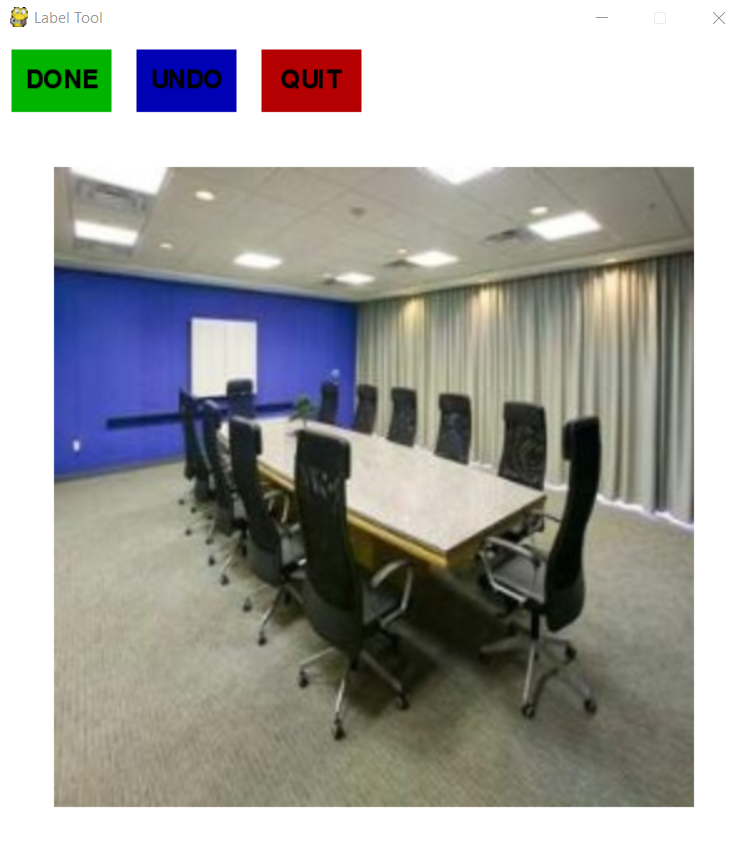
*Figure 7 – Clockwise from top left: classroom, bedroom, conference room, living room*

Data labelling was done by tracing structural edges in each image. However, popular online tools such as MIT LabelMe and Amazon Mechanical Turk did not have the capabilities suitable for this task, hence a custom labelling tool must be built. Using Python and the Pygame library, a simple labelling tool was designed (Figure 8).

Given that edges make up very few pixels in any given image, one concern is that the optimised model might get trapped in a local minimum and generate blank images. In order to prevent this, a larger penalty must be created. By creating a heatmap around each edge (Figure 9), it not only increases the penalty of generating a blank image, it also allows the model to learn when it is near the coordinates of an edge.

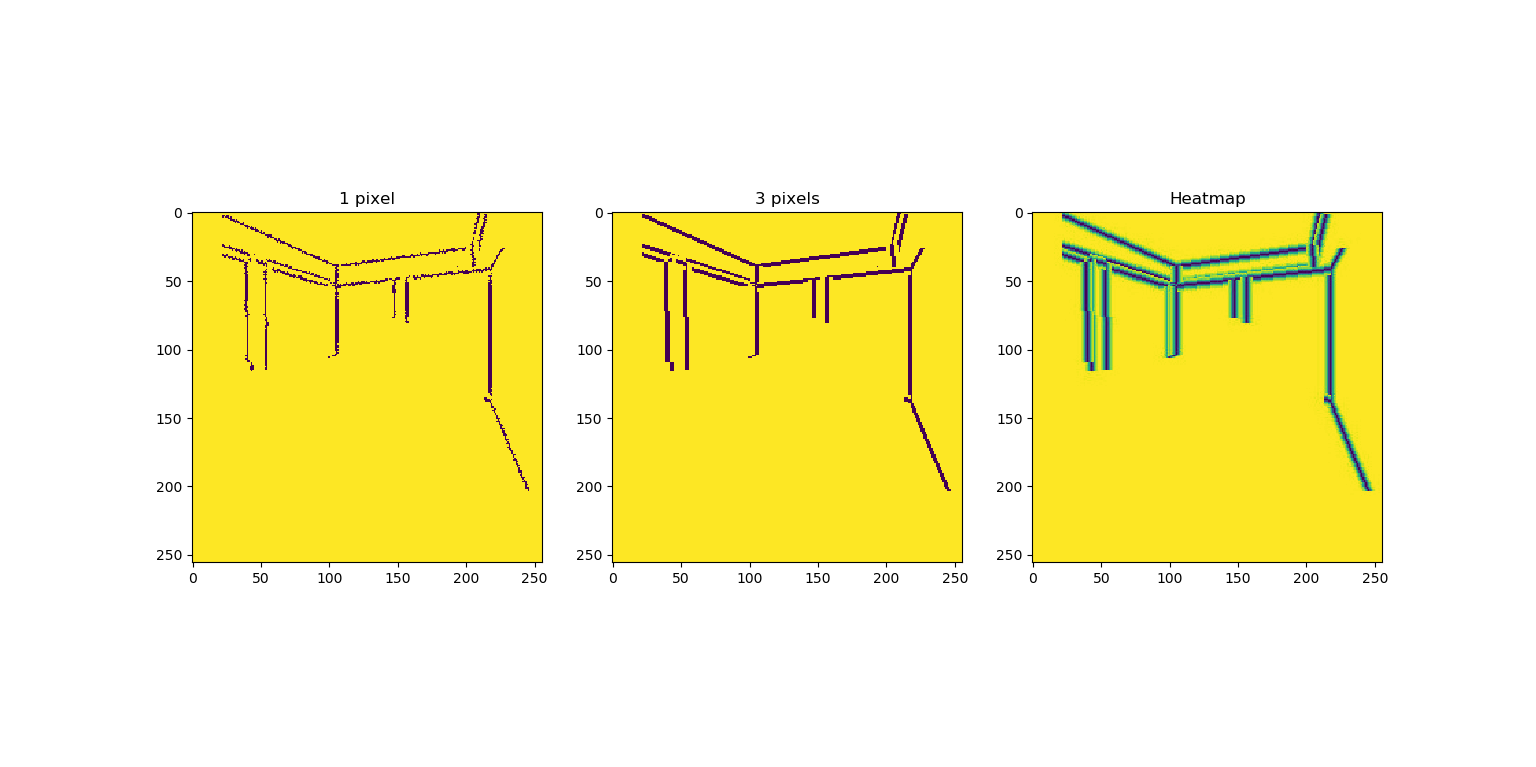


*Figure 9 – Example of Heatmap Feature Around an Edge*



*Figure 8 – Custom Label Tool*

Since the labelling tool was custom-built, this “heatmap” feature could be easily built into the tool, hence automating this process. A total of 3 different types of edges were used – 1 pixel edge, 3 pixel edge and heatmap edge (Figure 10).



*Figure 10 – Various edge types*

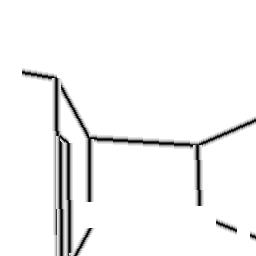
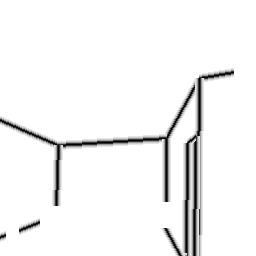
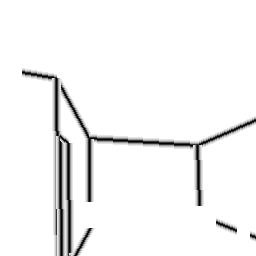
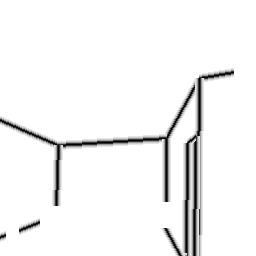
**4.3 Data Augmentation**

Research has shown that for image-related neural networks, the size of the dataset should be at least on the order of 104. In order to significantly increase the labelled dataset of 5,000 images, common data augmentation techniques for images were used. Using the Open Computer Vision (OpenCV) library, all images are subjected to three augmentations (Figure 11):

1. A horizontal flip
2. An increase in brightness
3. A horizontal flip plus an increase in brightness

The corresponding labels to the augmented image must also be altered (Figure 12):

1. A horizontal flip to match the flipped image
2. No change since brightness does not affect the location of structural edges
3. A horizontal flip to match the flipped image



*Figure 12 – Clockwise from top left: Original, Flipped, Brightened, Flipped + Brightened*



*Figure 11 – Clockwise from top left: Original, Flipped, Brightened, Flipped + Brightened*

Using the aforementioned augmentation techniques, a total of 20,000 labelled images is contained within the dataset. The new dataset is then further split in a 7:3 ratio to obtain a train set and a test set. The 14,000 labelled images in the train set is used to train the various neural networks, while the 6,000 labelled images in the test set is used to validate the neural network and prevent overfitting.

**4.4 Data Processing**

Each RGB input image to the neural network is then further normalized and standardized. This is common practice in training neural networks as it is beneficial to ensure that the entire dataset has a standard normal distribution (B, 2017). The detailed normalisation for each channel is as such:

Each channel’s mean and standard deviation is calculated based on the training dataset. Each image in the training dataset is assigned a value based on the mean value of its pixels. These values are then used to calculate mean and standard deviation of the entire dataset as shown below:

The final values obtained for each channel is given below:

**4.5 Training the Model**

For each neural network, a suitable loss function must be chosen to measure the accuracy of the model. It was initially assumed that Mean Squared Error Loss (MSE) could be used as a loss function. MSE measures the deviance of the predicted value of each pixel from the true value. However, the training failed to yield any meaningful results. Upon further investigation, it was determined that MSE was only suitable for linear regression models, whereas this project uses image segmentation and is therefore considered a classification model.

As such, a loss function that measures the accuracy of the classification of each pixel would be suitable. To that end, 2 loss functions may be relevant – Cross Entropy Loss and Binary Cross Entropy Loss. Cross Entropy Loss measures the performance of a classification model whose output is a probability value between 0 and 1 (Loss Functions, 2017). It is frequently used in multi-class classification problems. In PyTorch, this is implemented using the function CrossEntropyLoss and has the following equation (Docs: torch.nn, 2019):

Binary Cross Entropy Loss is a special case of Cross Entropy Loss where the classification problem is strictly binary. In the case of edge detection, this is relevant since each pixel is either classified as an edge or not an edge. In PyTorch, this is implemented using the function BCEWithLogitsLoss and has the following equation (Docs: torch.nn, 2019):

While both loss functions could theoretically work well for this classification problem, it was discovered that BCEWithLogitsLoss worked much better. The reason for this difference in performance is unknown.

The optimisation function that was chosen for this project was Adam. While there has been some concerns regarding its convergence (Bushaev, 2018), Adam and SGD remain the two most popular optimisers for deep neural networks. The models are trained at a learning rate of 0.001.

With these considerations, the three models based on SegNet, U-Net and DenseNet are trained on the train set for 50 epochs each. Each epoch attempts to fit each data in the train set exactly once. The model with the highest accuracy is then selected for further finetuning. The detailed results are highlighted in the next section.

**5. Results**

**5.1 Training results**

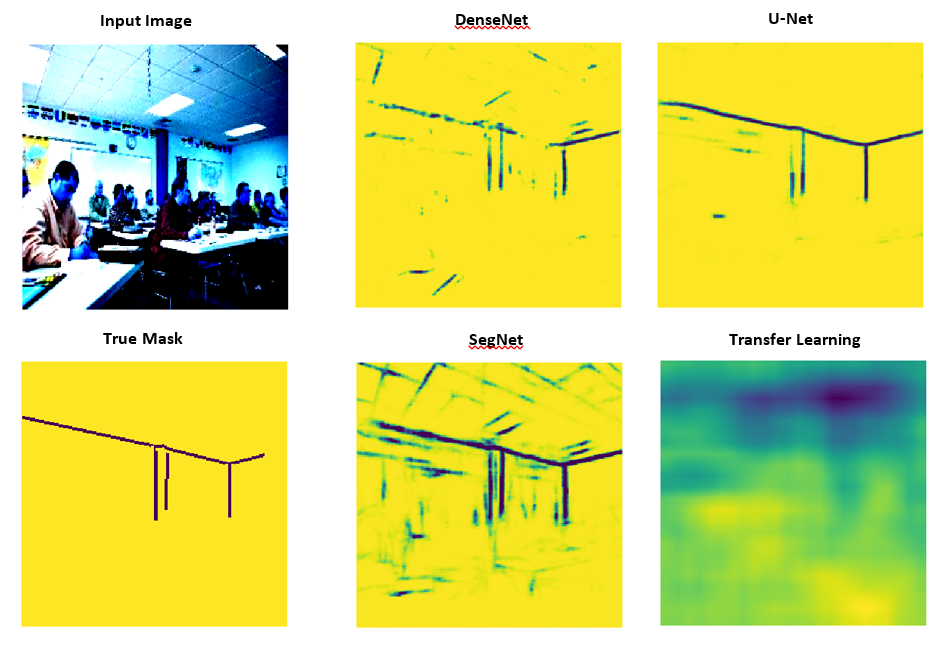
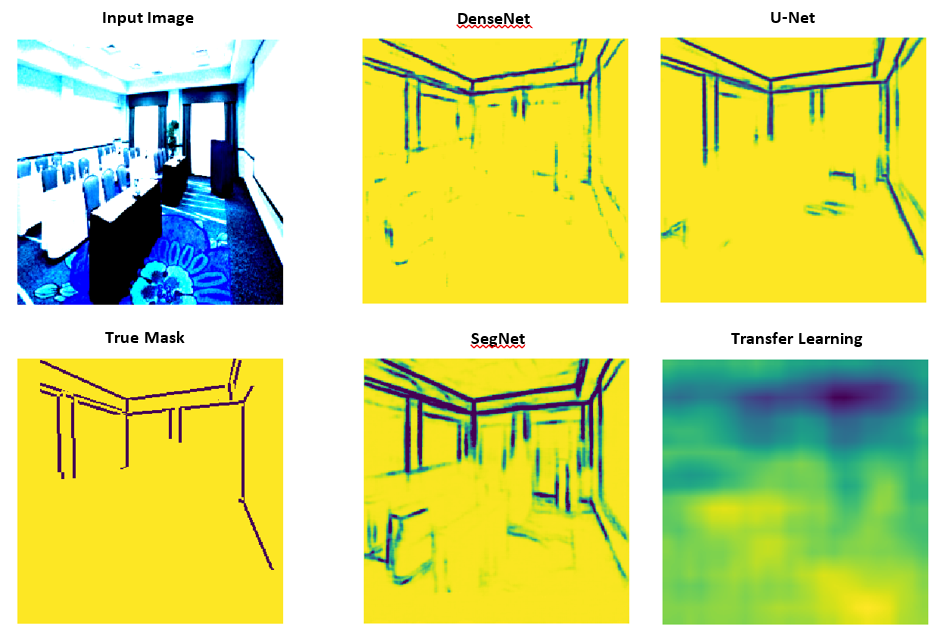
The three models based on SegNet, U-Net and DenseNet were each trained on 3 pixel wide edges over 50 epochs. The accuracy of the model is cross-validated with the test dataset and is summarised in Table 1. The loss is calculated using Binary Cross Entropy, while the accuracy is calculated using absolute error as shown in the following equation:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Loss** | **Accuracy** | **Parameters Size (MB)** |
| SegNet | 0.0562 | 0.9616 | 8.75 |
| U-Net | 0.0506 | 0.9683 | 29.38 |
| DenseNet | 0.0436 | 0.9740 | 20.14 |
| Transfer Learning | 0.0252 | 0.9998 | NIL |

*Table 1 – Summary of performance of various models*

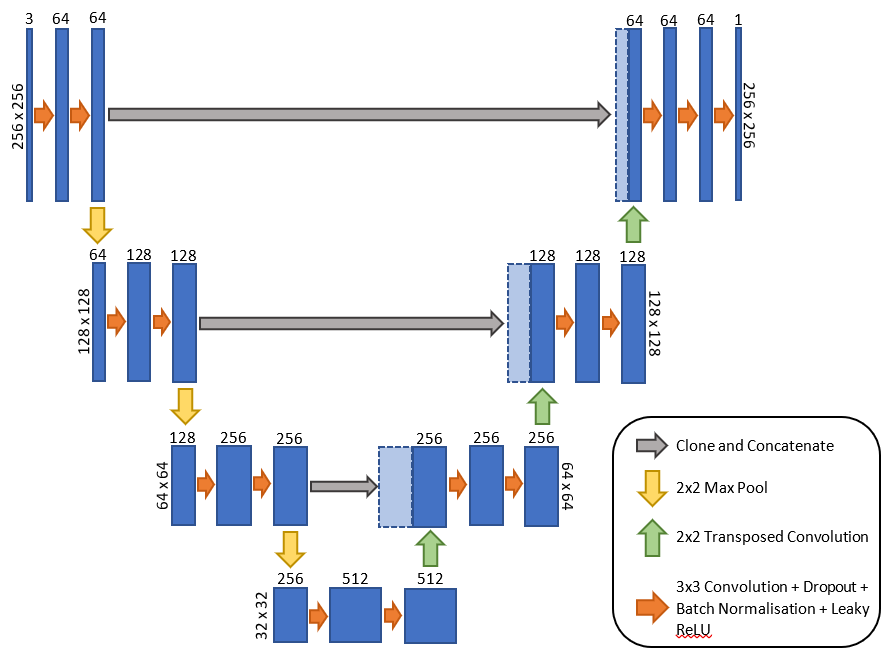
Evidently, SegNet is the most lightweight model, but it is also the least accurate. On the other hand, the transfer learning model based on FCN-RESNET101 scores the highest in terms of both loss and accuracy, while U-Net is the most memory-intensive. Based on these results alone, it would seem the transfer learning model has the best performance. However, closer inspection at the visual output from the various models indicates that models that work well on paper may not be the best. As shown in Figure 13, the output from the transfer learning Model does not contain any recognisable edges despite its stellar performance. In fact, U-Net has the best performance in identifying relevant structural edges.

*Figure 13 – Sample predicted masks from various models*



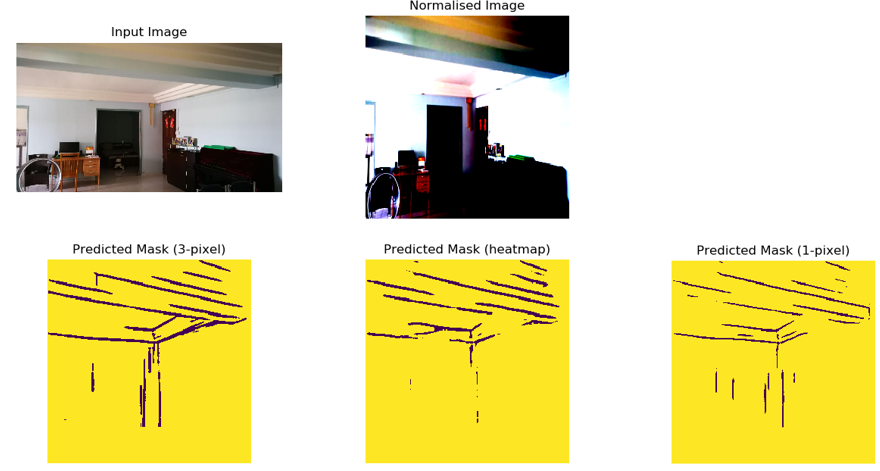
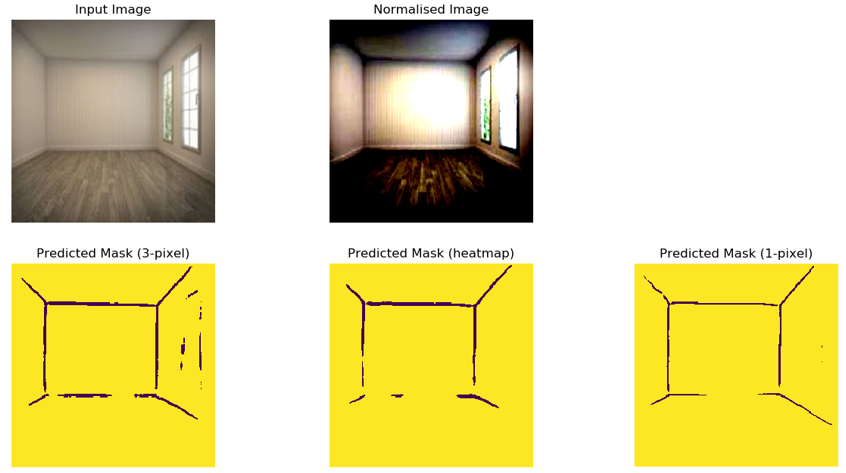
As such, the final model based on U-Net that was used is summarised in Figure 13:

*Figure 14 – Summary of model architecture*



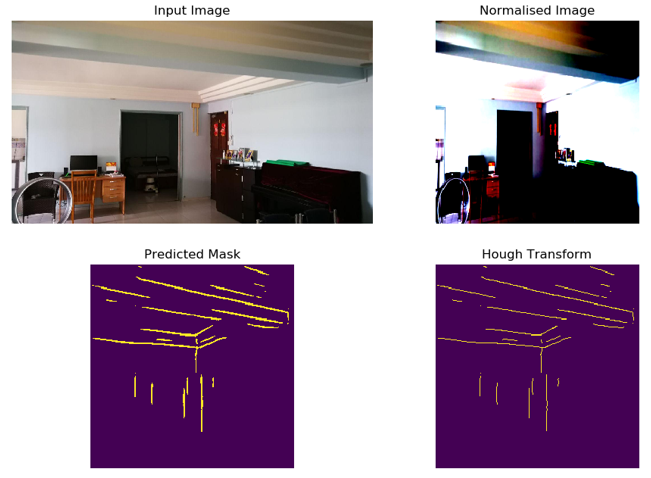
Further investigation was made into the effect of the type of edges used in the quality of the trained model. After training, the output of the model is processed using a threshold value of 0.5 to classify each pixel as either 0 or 1. Some sample results from testing the 1-pixel, 3-pixel and heatmap edges on the U-Net model is shown in Figure 15. From these results, it is determined that the model trained on 1-pixel edges has the best performance.

Using this information, the U-Net model is then retrained on the 1-pixel edge dataset over 50 epochs to attain the ideal number of epochs for the model. The training and testing losses are plotted against epochs to identify the epoch number where testing losses are minimal (Figure 16). This is to prevent any overfitting of the model on the trainset. From the results, the optimal number of epochs is 30.



*Figure 15 – Effect of edge type on model performance*

As shown in Figure 15, the output masks from the U-Net model that is trained on the 1-pixel edge dataset still had significantly thick edges. As such, the edges are first thinned using the Zhang-Suen thinning algorithm and is implemented using the OpenCV library. From there, the line equations of each edge are then obtained using a Probabilistic Hough Transform. The final output is shown in Figure 17.



*Figure 17 – Mask after thinning and Hough Transform*

**6. Conclusion**

**6.1 Summary**

The results derived from this project shows that machine learning, in particular Fully Convolutional Networks, can be used for detecting structural edges in photographic images and it is shown that it performs better than classical algorithms. The difference in performance is especially pronounced in photographs with large amounts of noise.

**6.2 Future Work**

Future work can be done to improve the mode by finetuning some of the hyperparameters such as the type of optimiser and the learning rate. In addition, the current model is not excellent at identifying floor edges since the dataset mostly contains images where the floor edges are hidden by objects. As such, there is a natural bias in the dataset which cannot be corrected easily. To rectify this, the dataset can be expanded in the future to include more empty rooms so that the model can be trained on a more balanced dataset. One way of doing it is to use a CAD software such as Solidworks to create digital renders of empty rooms.

Lastly, some edges in the photographs are partially hidden behind noise such as furniture and people. These edges would therefore be broken into smaller segments. As such, further work needs to be done in developing a robust algorithm that can extrapolate broken line segments to reconstruct the original line such that a wireframe model can be obtained.

# Bibliography

B, N. (2017, Sep 11). *Image Data Pre-Processing for Neural Networks*. Retrieved from Becoming Human: https://becominghuman.ai/image-data-pre-processing-for-neural-networks-498289068258

Bushaev, V. (2018, Oct 22). *Adam — latest trends in deep learning optimization.* Retrieved from Towards Data Science: https://towardsdatascience.com/adam-latest-trends-in-deep-learning-optimization-6be9a291375c

*Docs: torch.nn*. (2019). Retrieved from Pytorch: https://pytorch.org/docs/stable/nn.html

Fung, V. (2017, June 16). *An Overview of ResNet and its Variants*. Retrieved from Towards Data Science: https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035

Jonathan Long, E. S. (2014). *Fully Convolutional Networks for Semantic Segmentation.*

*Loss Functions*. (2017). Retrieved from Machine Learning Glossary: https://ml-cheatsheet.readthedocs.io/en/latest/loss\_functions.html

Marr, B. (2018, October 1). *Forbes*. Retrieved from What Is Deep Learning AI? A Simple Guide With 8 Practical Examples: https://www.forbes.com/sites/bernardmarr/2018/10/01/what-is-deep-learning-ai-a-simple-guide-with-8-practical-examples/#5881c2a08d4b

Olaf Ronneberger, P. F. (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation.*

Shah, M. (2017, June 1). *Semantic Segmentation using Fully Convolutional Networks over the years*. Retrieved from Github: https://meetshah1995.github.io/semantic-segmentation/deep-learning/pytorch/visdom/2017/06/01/semantic-segmentation-over-the-years.html

Simon Jégou, M. D. (2016). *The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation.*

Vijay Badrinarayanan, A. K. (2015). *SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation.*