Australasian Transport Research Forum 2023 Proceedings
29 November – 1 December, Perth, Australia
Publication website: http://www.atrf.info

Step by Step: Artificially intelligent models for predicting the footpath network using semantic segmentation

Jade Sams[1]

[1] SMEC; This research was completed in the fulfilment of the requirements for Honours in a Bachelor of Civil Engineering (Honours) at UNSW

Email for correspondence: jade.sams@smec.com

Abstract

Understanding the footpath network from the point of view of a pedestrian is often ignored by commercial mapping companies to the detriment of vulnerable footpath users, such as the visually impaired. Semantic segmentation is the process of assigning a categorical label to each pixel in an image (e.g. separating a picture of the urban landscape into the classes of footpath, roadway, sky and tree). The rise of autonomous vehicles has seen the creation of efficient artificially intelligent (AI) models to semantically segment the road network. Unfortunately, no equivalent model exists for understanding the footpath network. This places limitations on the navigation tools that connect vulnerable footpath-users with this much used piece of urban infrastructure, of particular importance for visually impaired pedestrians or for robot delivery services that interact with the footpath network in real-time.

This paper summarises research undertaken to semantically segment a world-first pedestrian-focused panoramic dataset (in collaboration with footpath.ai). A robust AI model, trained on this dataset, was then developed to accurately predict the Australian footpath network. Finally, an overview is presented of two robust and efficient models with a real-time prediction time of 22 milliseconds and an overall accuracy of 95% during training.

The key finding of this study was that the type of image that the model will be tested on must be included in what the model is trained on; here the model was tested on panoramic images, therefore in this instance a robust model required panoramic images to be included in the training dataset. The opportunities and limitations associated with the use of models going forward are also discussed.

1. Introduction

Technological advances in visual object detection methods and the increasing demand for the mapping of the road network saw Google Maps expand their origin-destination services in the late 2000s and launch Google Street View. Today, Google has mapped the road network in 83 countries and has made this data readily available for a global audience. However, no such application exists for the mapping of other urban infrastructures, more specifically the footpath network. The rise of autonomous vehicle production has seen companies, such as Tesla, train artificial intelligence (AI) systems to identify its surroundings using semantic segmentation.

This study, in association with the Sydney based start-up footpath.ai, will investigate the mapping and understanding of Australia's footpath network using artificial intelligence software, trained through semantic segmentation. The aim of this research is to produce a

robust model which can be used for real-time semantic segmentation of the footpath network, and such a project will have wide-reaching impacts for vision-impaired and mobility-impaired pedestrians.

2. Literature review

2.1. Semantic segmentation

Segmentation is the process of deconstructing a digital image into different regions which all share similar characteristics (Halder & Pramanik 2012). Segmentation predicts detailed masks over a region of an image, as opposed to the bounding boxes predicted by object detection methods. Semantic segmentation is the computational task of assigning categorical labels to each pixel in an image (Orhan & Bastanlar 2021). A segmentation mask is drawn over an object by the annotator by outlining the object's edges and automatically filling the space between. During the annotation process, the annotator may find it easier to overlap multiple masks when annotating obscured objects (such as a building that is situated behind a tree). However, the final image will always only allow each pixel to be assigned to one class or sub-category, therefore the annotator must be precise while layering their semantic masks.

Erfani et al. (2022) used the free, open source, web-based tool CVAT to annotate images. This tool allows the user to manually define the boundaries of an object's mask using a computer mouse. Additionally, the tool allows the user to define an unlimited number of categories to segment an image. For efficiency and to promote the availability of future model extensions, it is noted in Orhan and Bastanlar (2021) that all new models should follow a similar class definition. As the Cityscapes dataset, created by Cordts et al. (2016), is currently the most challenging dataset (Acuna et al. 2018, p. 5), due to its comprehensive dataset of 5,000 semantically labelled images and diversity of weather conditions, it would be ideal for new models to base their class definitions off this work to complement each other.

2.2. Model training networks

Once a dataset has manually annotated images, they will be used to train a network which in turn creates a model that can automatically annotate other images. The better trained the model is, the better its accuracy will be when it is used in the future. Historically, neural networks were used as deep learning technology had not yet been developed. Following important advancements in deep learning technology, the field of segmentation models began using deep neural networks (DNNs). The most prevalent DNNs can be classified as convolutional neural networks (CNNs) or generative adversarial networks (GANs). Popular CNNs used for real-time semantic segmentation in literature are U-Net (Guo et al. 2018), ERFNet (Romera et al. 2017) and its extension ERF-PSPNet (Yang et al. 2020), and ICNet (Zhao et al. 2018).

2.3. Annotated datasets

The availability of diverse image datasets is one of the most challenging factors in this field, as a supervised network cannot be trained without an adequate number of images collected and semantically annotated. Additionally, for a model to be evaluated and used operationally, the image dataset must be wide enough to include training images, validation images, and testing images. Specifically in the footpath network domain, Venkatesh et al. (2021) noted that most available datasets were captured from the viewpoint of a vehicle as opposed to the pedestrian. Due to the time and cost required to collect such data, some authors have instead opted to create synthetic images to train their network, however this decision reduces the accuracy of the final model when real-world images are introduced during the testing stages (Romera et al. 2018, p.

1829). Some of the most common annotated datasets in the field and their applications are discussed below.

2.3.1. KITTI Imagery

Geiger et al. (2012) responded to the lack of real-world benchmarks to be used in visual recognition systems through the creation of KITTI Imagery. The team captured images for robotic application from cameras attached to the top of a car which travelled in rural areas and along highways in Germany. Rather than relying upon online crowd-sourced annotations, the team hired annotators specifically for this project. The dataset includes 12,000 images whose annotations have been broken down into 16 classes. The project focused upon 3D object detection; hence 3D bounding boxes were used. This analysis found more than 40,000 objects in the 12,000 images, highlighting the dataset's comprehensive nature.

2.3.2. Cityscapes

Cityscapes was created by Cordts et al. (2016) in response to the lack of comprehensive datasets with diverse (or challenging) semantically segmented urban scenes. The aim of the authors was to create a benchmark tailored for autonomous driving research. The dataset includes 5,000 fine pixel annotated images and 20,000 coarsely annotated images where only significant clusters of pixels have been labelled. All images are of equirectangular nature. Cityscapes is composed of video sequences recorded from 50 cities in Germany, which are then extracted as images, and covers differing weather conditions (however no adverse conditions were included). The dataset defined 30 annotation classes which are grouped into eight categories. A major contribution of this work to the field was the public release of its semantically segmented data including 2,975 images for training, 500 images for validation, and 1,525 images for testing (Acuna et al. 2018).

2.3.3. SYNTHIA-PANO

The synthetic dataset of SYNTHIA-PANO was created by Xu et al. (2019) due to the scene understanding advantages offered by panoramic images, however, the field lacked real-world panoramic datasets at the time of publishing. The larger field-of-view was seen to offer greater information capacity and scene stability for models to act more robustly against image distortion than previous equirectangular images could offer. The SYNTHIA-PANO dataset contains 16 classes for label annotation. Images were created to reflect seasonal differences in summer and autumn. All synthetic images were stitched as a panorama from the SYNTHIA dataset which includes four-directional images, created from computer-rendered 3D city traffic scenes in New York and a generalised European town. 1,800 images exist for the New York-like city, and 1,430 images exist for the synthetic European town. However, even though simulated image technology is improving to match realistic images, Romera et al. (2018) found that deep models trained only on synthetic datasets perform poorly when real world images are introduced. Therefore, synthetic panoramic imagery is best used in practice as an additional training set for networks when fused with real-world semantically segmented images.

2.3.4. Panoramic Annular Semantic Segmentation (PASS)

Yang et al. (2019) utilised the traditional field-of-view semantically segmented datasets of the Chinese streetscape and fused them together to create a smooth yet synthetic panoramic scene. Using these panoramas, this paper proposed a model which they named the Panoramic Annular Semantic Segmentation (PASS) which was trained using the ERF-PSPNet framework. The PASS dataset contains 400 finely labelled panoramas with only 6 critical classes: Car, Road, Crosswalk, Curb, Person, and Sidewalk. As such, the generalisation abilities of this dataset are limited due to the vast number of pixels that are assigned the label of 'Other'.

2.3.5. Crowd4Access

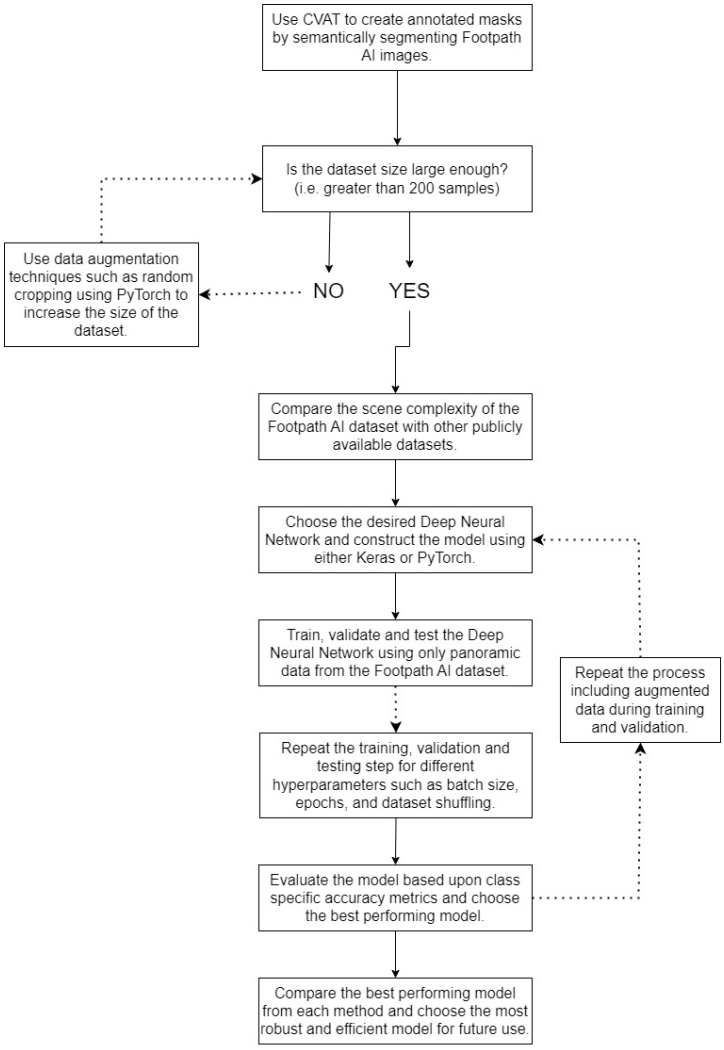
The accessibility of the footpath for users with mobility issues has been recently investigated by the team of Venkatesh et al. (2021). The team noticed the lack of annotated datasets that were not from the perspective of a moving vehicle; the same gap that saw the creation of footpath.ai. As such, the team crowdsourced urban images around Ireland using the phone application Mapillary and created an open-source dataset. Although the team acquired 39,642 images in differing weather and time of day conditions, it is unlikely that these can be used for footpath.ai training as they do not capture the required panoramic context. Additionally, Crowd4Access has not completed its manual semantic segmentation stage as of the publish date of this paper. The team instead initially focused on the object detection of tactile pavements, hence only semantically segmenting these instances; a sub-class of footpath.ai's semantic segmentation. Venkatesh et al. (2021) trained an AI model to semantically segment its dataset with moderate success on the Cityscapes dataset using ICNet. As a result, no ground truth labels currently exist for this footpath dataset.

3. Research methodology

This research project focused upon the creation of an efficient artificially intelligent model to map the footpath network; hence, a comprehensive pedestrian view dataset was manually obtained by the footpath.ai team and semantically segmented. Machine learning frameworks such as Keras offered by the programming language Python were utilised to train a deep learning model. This section offers an overview of the collection of footpath.ai's own dataset and an introduction to the two deep learning frameworks.

An outline of the methodology to train a deep learning model can be found in Figure 1 below.

Figure 1: Methodology flowchart to create an efficient AI model using the footpath.ai dataset



3.1. footpath.ai dataset

After analysing the collected data, the footpath.ai team determined that there were 47 common classes found in the urban landscape images. Table 1 below summarises the 48 classes adopted, with one class titled 'Other' being chosen when the object did not fit neatly into any of the classes.

Table 1: The 48 segmentation classes footpath.ai used during annotation, with their background colours reflecting their manually assigned RGB value

Footpath	Building	Pet	Traffic signal	Bin	Bike rack
Bike lane	Wall	Ramp	Pole	Advertising board	Tree
Car lane	Platform	Stairs	Sign	Tree Surrounding - Protector	Bush
Pedestrian Crossing	Footpath canopy	Motorbike	Traffic cone	Electric box	Grass
Tactile Paving	Fence	Bicycle	Barricade	Traffic control box	Sky

Tree Surrounding - Permeable Paving	Pedestrian	Scooter/ wheelchair	Bench	Post Box	Water
Driveway	Cyclist	Small Vehicle	Café Chair	Firehose	Terrain
Railway Track	Bird	Large Vehicle	Café Table	Bus/tram shelter	Other

Of these 48 classes, seven were determined to be of significant importance due to their high occurrence across all images and/or their interaction with the footpath. All other classes were combined into an eight 'Other' category. These classes were chosen to form the baseline for the initial model calibration. The classes (in no particular order) were:

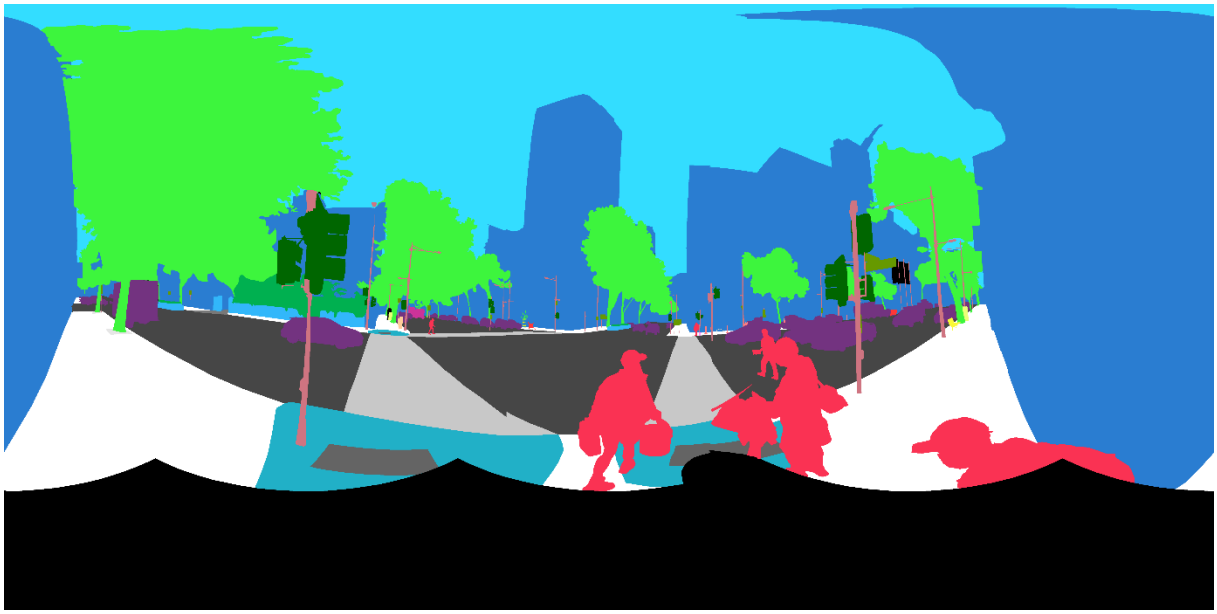
Footpath; Pedestrian Crossing; Car Lane; Building; Wall; Tree; Sky.

'Building' and 'Wall' were chosen as they bound the footpath; 'Wall' is a subset of a 'Building' and is distinguished by having an awning covering the building, commonly found in a shopfront region.

An example of a ground truth image and its annotated mask in CVAT of a scenic image in Victoria from the footpath.ai dataset are seen in Figure 2 and Figure 3 respectively.

Figure 2: Ground truth panoramic image from the footpath.ai dataset

Figure 3: Semantically segmented annotation mask completed in CVAT using 48 classes

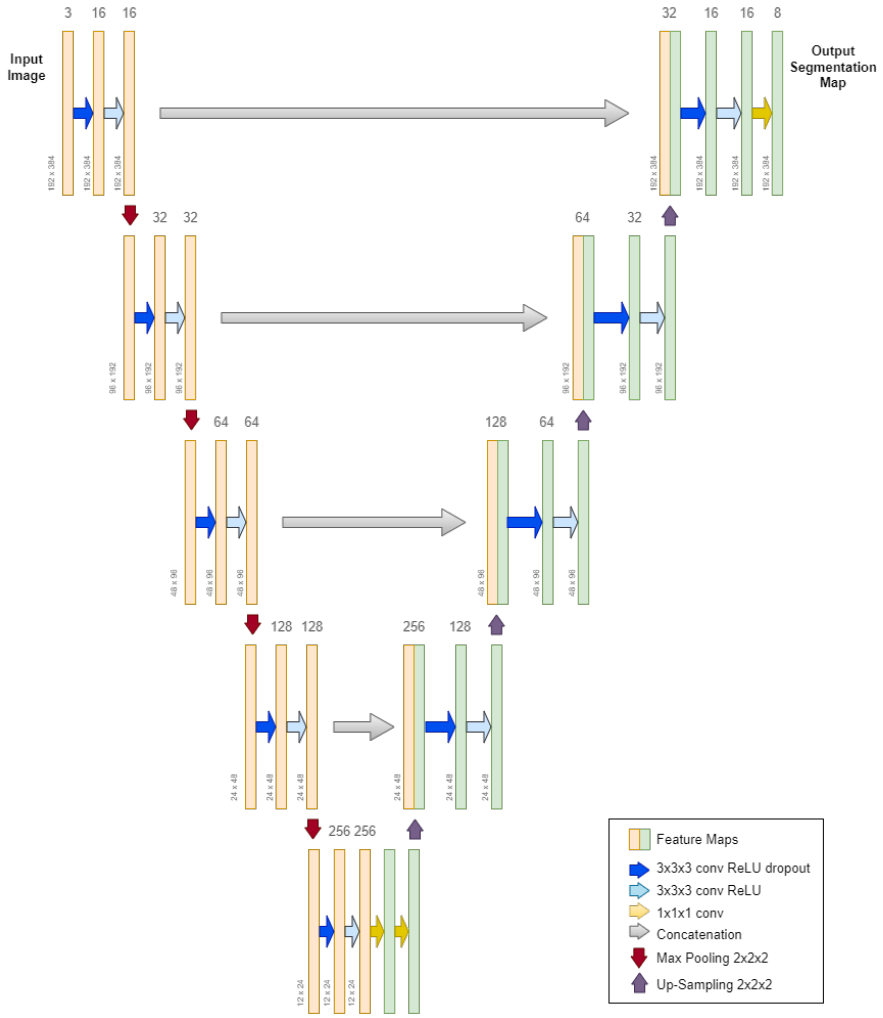


The footpath.ai dataset currently has 98 completed annotation masks due to the hours it takes to complete one annotation and the small team size. Therefore, to improve model results the data augmentation technique of random cropping of the image and its corresponding annotated mask was used to increase the size of the dataset. This cropping technique is beneficial as the image does not lose its proportions and the objects within are not shrunk or enlarged to be a different scale than reality. A key advantage of using panoramic images as the training dataset is that the model trains on a wider field-of-view which holds greater information capacity and scene stability. These cropped results still hold the desired scene complexity yet offer multiple angles of the same scene. Thus, with the training set size of the model will increase, the model performance may also increase at a fraction of the time it takes to annotate further panoramic images.

3.2. Model construction

Based upon a literature review, U-Net was chosen as the initial model to be built and trained by the footpath.ai team. U-Net is a popular framework for other outdoor panoramic semantic segmentation projects, and as a consequence there are many online resources for its implementation on new datasets. The architecture that was implemented in this study has been replicated in Figure 4 below.

Figure 4: U-Net Architecture that has been implemented in this paper. The left side represents the contracting path, and the right side represents the expanding path



The following hyperparameters of the model were modified during analysis to determine the combination that creates the best performing model both qualitatively and quantitatively.

Batch Size: The batch size of a model refers to the number of samples in the training set the model will train on before it updates its parameters. For example, with a training set of 100 images and a batch size of 5, the model will update its parameters 20 times in each run. The smaller the batch size, the more times the model will update its parameters, and theoretically the better it will perform.

Epochs: The number of epochs of a model is the number of times the model will run during its training phase. The number of epochs along with the batch size will determine how many iterations the model will perform which may be very computationally heavy. Using the previous example, for a training set of 100 images, a batch size of 5 meaning 20 updates per

run, and using 200 epochs, the number of iterations will be 20 multiplied by 200 totalling 4,000 iterations. While it may seem logical that the more epochs a model runs the better its performance, graphical representations of the model's loss may reveal that a global minimum or asymptote is reached at a certain point.

Shuffle: Shuffle refers to the use of shuffling when choosing a subset of images in the model's training dataset for each batch. This option can either be set to True or False. A shuffled dataset is ideal where the dataset has clusters of similar scenery.

3.3. Model verification metrics

3.3.1. Accuracy

The pixel accuracy of a network's output is determined through a comparison with the manual semantically segmented ground truth (GT) reference. An average of per-class accuracy as viewed in Equation 1 is a good metric for datasets that only have a small number of pixels belonging to a 'void' or 'other' category (Fooladgar & Kasaei 2019). Here, C represents the number of classes, n represents the number of pixels with the class label i in the output, and t is the total number of pixels belonging to class i in the GT.

$$\text{mean}(Pixel_{acc}) = \frac{1}{C}\sum_{i=1}^{C}\frac{n_{ii}}{t_i} \tag{1}$$

3.3.2. Jaccard Index

The standard evaluation metric is known as the Jaccard Index, or more commonly the intersection over union (IoU) metric (Cordts et al. 2016). This is a similarity measure between the output's predicted labels and the GT reference labels.

The general mean intersection over union (mIoU) is more common as it can be broken down for each category or class. Pixels that are labelled as 'other' will not be counted in this metric. In Equation 2 below TP, FP and FN stand for true positive, false positive, and false negative pixels, respectively.

$$mIoU = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FP_i + FN_i} \tag{2}$$

A true positive occurs when the predicted pixel correctly matches the ground truth pixel. A false positive occurs when the model incorrectly predicts the pixel as belonging to the selected class which disagrees with the ground truth which shows it is a part of a different class. A false negative occurs when the model predicts that the pixel is not a part of the selected class, however the ground truth confirms that it should be.

3.3.3. Multi-class normalisation

While the creation of a highly accurate model is important, this value is only a useful qualifier if it is not skewed by classes of lesser importance. For the footpath.ai team, the classes that are associated with the 'Footpath' and its surrounding objects, such as 'Pedestrian Crossings' and a 'Car Lane', are of the utmost importance when it comes to choosing the most appropriate final model. For example, a particular set of hyperparameters may achieve an accuracy score higher than 80%, however if its good performance is solely based upon its highly accurate prediction of the 'Other', 'Sky' and 'Tree' classes, this is likely not be utilised for the purposes of this study. As such, a normalisation process will occur. The weights to be awarded to each class are listed below in Table 2.

Table 2: List of normalised weights to be attached to each class during evaluation to find the normalised mIoU

Class	Normalised Weight
Footpath	0.25
Pedestrian Crossing	0.2
Car Lane	0.2
Building	0.1
Wall	0.1
Tree	0.05
Sky	0.05
Other	0.05
Total	**1**

4. Results analysis

Five methods were tested, and their results were analysed, to determine the best performing model based upon prediction accuracy and the relevance of model results to footpath understanding. The five methods were:

1. Base Case.
2. Panoramas for both training and testing.
3. Augmented data for training and Panoramas for testing.
4. Augmented data (all) & Panoramas for training and Panoramas for testing.
5. Augmented data (subset) & Panoramas for training and Panoramas for testing.

4.1. Base case

The baseline hyperparameters for Method 1 were randomly chosen to be a batch size of 16, 100 epochs, and setting the model's shuffle function to False. The verification metrics that were outputted for each set were their Jaccard Index (ideally close to 1), their loss function value (ideally close to 0), and accuracy (ideally close to 1).

4.2. Panoramas for training and testing

Method 2 involved adjusting the hyperparameters randomly assigned to the Base Case to determine the most efficient combination. The hyperparameters that were modified were the batch size, the number of epochs to be completed, and whether the dataset was shuffled or not. This paper compared batch sizes of 1, 4, 16 and 32 and epochs of 100 and 200. The number of model iterations is dependent upon the size of the training set, the batch size and the number of epochs run. It is hypothesised that the more iterations that occur the better the model will perform.

The results of the best performing models saw a batch size of 16, with 200 epochs and shuffle set to false perform better than the base case, with a training accuracy of 95% and a normalised mIoU of 45%.

The output predictions from the Base Case and Method 2 qualitatively highlight that the model struggled to create harsh boundaries between the footpath and the car lane, and also appeared to incorrectly classify pixels that should be walls as buildings. When reviewing the original

images in the dataset, it is apparent that a pedestrian crossing, while always found on the road, is sometimes occluded by the sunlight making it harder for the model to distinguish between the two classes. In a similar vein, the building and wall pixels look similar without knowing the distinction between the classes - a 'Wall' is considered the canopy of any building covering the footpath. An additional pixel-class analysis reveals that in the current footpath.ai dataset 1% of pixels are classified as 'Pedestrian Crossing' compared to the 9% of 'Car Lane', and only 10% of pixels are classed as 'Wall' compared to the 31% labelled 'Building'. To overcome this obstacle, this paper hypothesises that the inclusion of geometric augmentations such as scaling and cropping the panoramas to focus upon only the 'Pedestrian Crossing' and 'Wall', as suggested by Romera et al. (2018), would increase the size of the training dataset in these classes and hence improve the model's predictions.

4.3. Data augmentation

The footpath.ai dataset is one of the first of its kind to be captured from the perspective of a pedestrian on the footpath. As a result, the size of the dataset is quite small compared to other publicly available datasets such as Cityscapes. To increase the size of the available dataset by 400% for testing and training purposes, three randomly cropped sections of each panorama were extracted.

Since 360-degree panoramic images are anticipated to be used in the real-world application of the footpath.ai model, it was deemed appropriate to not include the augmented images in the testing dataset. Therefore, two scenarios were tested, and their results are set out below. The two scenarios were:

1. Using the augmented data only for the training and validation of the model and using the panoramic data for testing.
2. Using both the augmented data and panoramic data for training and validation of the model and using only the panoramic data for testing.

Results of the first scenario, Method 3, indicated that training a model on solely augmented data but then testing it on panoramic data elicits an overall poor performance of the model, with the normalised mIoU of testing reaching only 28%. Therefore, both augmented and panoramic data were combined for future training. As a result, the shuffle hyperparameter will always be set to True to force dataset variety during training. However, since the augmented dataset is triple the size of the available panoramic dataset there is a chance that the model will favour the augmented data over the panoramas during training. Therefore, this scenario was split up into two sub-categories to test whether the number of augmented images in the training dataset affected the model's performance:

- Method 4 involved all available augmented and panoramic images in the training dataset.
- A selection of augmented images in the 'Pedestrian Crossing' and 'Wall' classes as these are the two worst performing classes during panorama training. This subset was combined with all panoramas for training and validation. The most efficient number of augmented images to use was unknown; further analysis determined that the model achieved the highest normalised mIoU when 24 augmented images were included during model training and validation, hence this selection was used for Method 5.

4.4. Comparison of best performing models

A summary of the verification metrics for the best performing models in each of the 5 methods are below in Table 3, and each of their respective class evaluation statistics are found in Table 4 for a comparison between the methods.

Training Set

Table 3: U-Net Verification Metrics for all the best performing scenarios in the tested methods on the Training Set. Cells highlighted in green represent the best performing method for each metric

Method	1	2	3	4	5
Batch Size	16	16	1	4	1
Epoch	100	200	200	100	200
Shuffle	False	False	False	True	True
Jaccard Index	0.65	0.86	0.70	0.45	0.85
Loss	0.40	0.13	0.35	0.82	0.14
Accuracy	0.85	0.95	0.88	0.72	0.95

Test Set

Table 4: Class Specific Jaccard Index of all the best performing scenarios in the tested methods on the footpath.ai dataset. Cells highlighted in green represent the best performing method for each class

Method	1	2	3	4	5
Batch Size	16	16	1	4	1
Epoch	100	200	200	100	200
Shuffle	False	False	False	True	True
Footpath	52%	59%	27%	18%	30%
Pedestrian Crossing	0%	16%	0.2%	13%	48%
Car Lane	37%	44%	34%	52%	41%
Building	52%	55%	41%	70%	75%
Wall	19%	19%	6%	10%	35%
Tree	44%	52%	45%	63%	64%
Sky	82%	88%	88%	92%	91%
Other	69%	69%	62%	38%	44%
mIoU	54%	62%	46%	38%	50%
Normalised mIoU	37%	45%	28%	35%	46%

A comparative analysis of the above results highlights that Method 2, where only footpath.ai panoramas were used for both training and testing, and Method 5, where footpath.ai panoramas and a subset of augmented images were used for training, attained the best results during the calibration process and during testing. Additionally, it is noted that the best performing models of Method 2 (see Figure 5) and Method 5 (see Figure 6) achieve a very similar overall performance for the training, validation, and test sets, and obtain a normalised mIoU of 45% and 46% respectively.

Figure 5: Model predictions of the best model from Method 2: batch size of 16 over 200 epochs with shuffle off

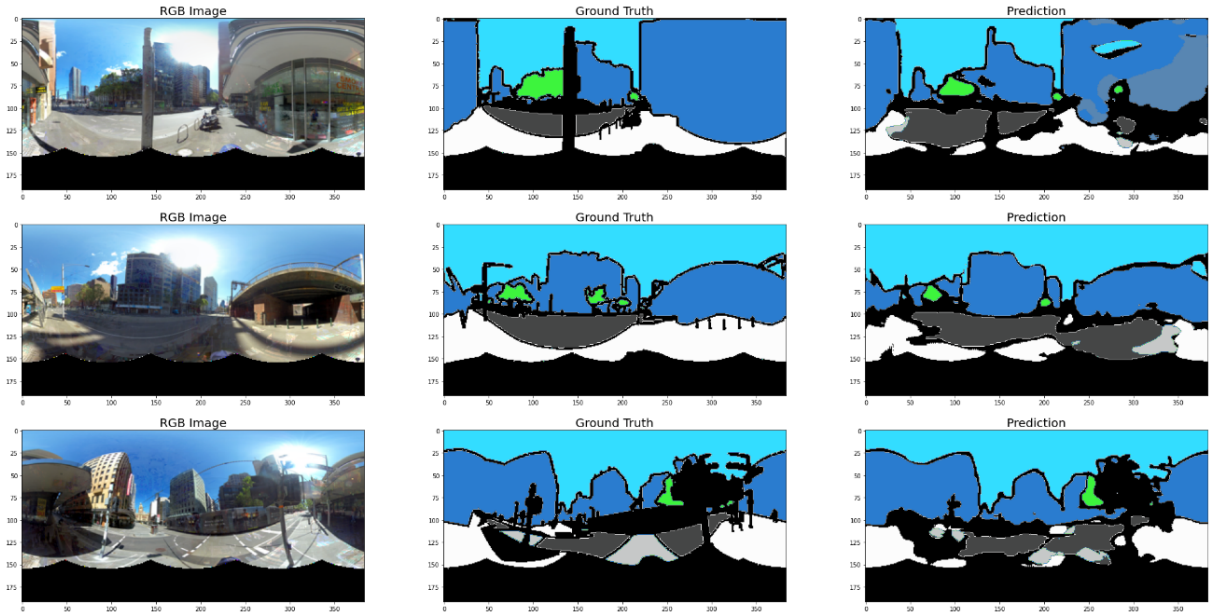
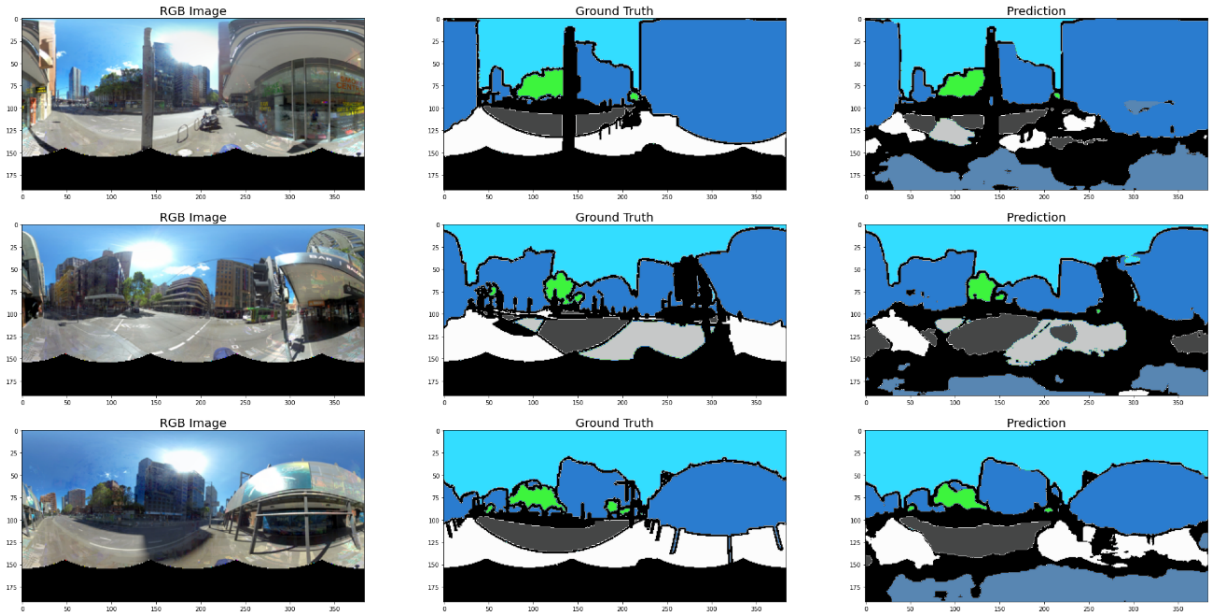


Figure 6: Model predictions of the best model from Method 5: batch size of 1 over 200 epochs with shuffle on



The distinction between these models lies within their class specific performance; Method 5 performs better than Method 2 in every class except for the 'Footpath' and 'Other', with a marked increase in the prediction of the 'Pedestrian Crossing' category. This outcome supports the introduction of a selection of augmented images into the training and validation datasets to improve the model's predictions for a specific class. Therefore the Method 5 model is suitable for implementation in a navigational tool for vulnerable visually-impaired pedestrians to accurately predict their walking route. On the contrary, if the model is to be used for the footpath network, and pedestrian crossings are not as important during operation, Method 2 would be the preferred model for implementation as it performs with a greater mIoU score of

59% in the 'Footpath' class compared to Method 5's 30%. The Method 2 model is suitable for robot delivery service applications that interact with the footpath in real-time.

Both models perform with a prediction time of 22 milliseconds when utilising GPUs and achieved metrics of 95% accuracy on the training dataset, 77% accuracy on the validation dataset, and scored a normalised mIoU of 46% during testing. Therefore, an efficient and robust model is created when the training dataset includes the style of imagery, here panoramas, that it will be tested on.

5. Conclusions

The two artificially intelligent models developed in this paper were successful in predicting and understanding the footpath scene in real-time. The models were trained, validated, and tested on the new footpath.ai dataset, also developed for this paper, a world first semantically segmented panoramic dataset from the view of the pedestrian. One of the models is more efficient at predicting the footpath itself which can be used for robot delivery service applications, while the other is more robust at understanding the objects connected to the footpath network such as pedestrian crossings and buildings, ideal for vulnerable visually impaired pedestrians to be applied to a navigation tool.

The footpath.ai dataset semantically segmented for this paper is particularly important for the transport field, as it initiates a shift for semantic segmentation studies to focus globally on modes of transport other than automobiles that use the road network. The scene complexity provided by the footpath.ai panoramas also set a benchmark for future works to compare their models against or, alternatively, extend upon by using a similar semantic class structure.

This work revealed that when training an artificially intelligent model for semantic segmentation scene understanding, the image form that the model will be tested on must be included in what it is trained on. Panoramic imagery was used in this study, however if in the real-world application of these models a smaller field-of-view is required, the models should be re-run following this paper's methodology.

6. Acknowledgments

I would like to give a special thanks to Dr. Meead Saberi, and the team members of footpath.ai, especially Jack Wang and Nouna Khandan, without whom this research would not have been possible.

7. Bibliography

Acuna D, Ling H, Kar A and Fidler S (2018) 'Efficient Interactive Annotation of Segmentation Datasets with Polygon-RNN++', *arXiv:1803.09693 [cs]* [Preprint], accessed: 15 March 2022. http://arxiv.org/abs/1803.09693

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S and Schiele B (2016) 'The Cityscapes Dataset for Semantic Urban Scene Understanding', *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, pp. 3213–3223, doi:10.1109/CVPR.2016.350

Erfani SMH, Wu Z, Wu X, Wang S and Goharian E (2022) 'ATLANTIS: A benchmark for semantic segmentation of waterbody images', *Environmental Modelling & Software*, 149, p. 105333, doi:10.1016/j.envsoft.2022.105333

Fooladgar F and Kasaei S (2019) 'A survey on indoor RGB-D semantic segmentation: from hand-crafted features to deep convolutional neural networks', *Multimedia Tools and Applications*, 79(7–8), pp. 4499–4524, doi:10.1007/s11042-019-7684-3

Geiger A, Lenz P and Urtasun R (2012) 'Are we ready for autonomous driving? The KITTI vision benchmark suite', *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI: IEEE, pp. 3354–3361, doi:10.1109/CVPR.2012.6248074

Guo Z, Shengoku H, Wu G, Chen Q, Yuan W, Shi X, Shao X, Xu Y and Shibasaki R (2018) 'Semantic Segmentation for Urban Planning Maps Based on U-Net', *2018 IEEE International Geoscience and Remote Sensing Symposium*, Valencia: IEEE, pp. 6187–6190, doi:10.1109/IGARSS.2018.8519049

Halder A and Pramanik S (2012) 'An Unsupervised Dynamic Image Segmentation using Fuzzy Hopfield Neural Network based Genetic Algorithm', *arXiv:1205.6572 [cs]* [Preprint], accessed: 3 March 2022. http://arxiv.org/abs/1205.6572

Orhan S and Bastanlar Y (2021) 'Semantic segmentation of outdoor panoramic images', *Signal, Image and Video Processing*, 16(3), pp. 643–650, doi:10.1007/s11760-021-02003-3

Romera E, Alvarez JM, Bergasa LM and Arroyo R (2017) 'Efficient ConvNet for real-time semantic segmentation', *2017 IEEE Intelligent Vehicles Symposium (IV)*, Los Angeles, CA, USA: IEEE, pp. 1789–1794, doi:10.1109/IVS.2017.7995966

Romera E, Bergasa LM, Alvarez JM and Trivedi M (2018) 'Train Here, Deploy There: Robust Segmentation in Unseen Domains', *2018 IEEE Intelligent Vehicles Symposium (IV)*, Changshu: IEEE, pp. 1828–1833, doi:10.1109/IVS.2018.8500561

Xu Y, Wang K, Yang K, Sun D and Fu J (2019) 'Semantic segmentation of panoramic images using a synthetic dataset', *Artificial Intelligence and Machine Learning in Defense Applications*, Strasbourg, France: SPIE, p. 9, doi:10.1117/12.2532494

Yang K, Hu X, Bergasa L, Romera E and Wang K (2019) 'PASS: Panoramic Annular Semantic Segmentation', *IEEE Transactions on Intelligent Transportation Systems*, 21(10), pp. 4171–4185, doi:10.1109/TITS.2019.2938965

Yang K, Hu X, Fang Y, Wang K and Stiefelhagen R (2020) 'Omnisupervised Omnidirectional Semantic Segmentation', *IEEE Transactions on Intelligent Transportation Systems*, 23(2), pp. 1184–1199, doi:10.1109/TITS.2020.3023331

Zhao H, Qi X, Shen X, Shi J and Jia J (2018) 'ICNet for Real-Time Semantic Segmentation on High-Resolution Images', *arXiv:1704.08545 [cs]* [Preprint], accessed 7 March 2022. http://arxiv.org/abs/1704.08545