land

MDPI

Article

Pre-Processing Training Data Improves Accuracy and Generalisability of Convolutional Neural Network Based Landscape Semantic Segmentation

Andrew Clark [1,2,*], Stuart Phinn [1] and Peter Scarth [1]

[1] Remote Sensing Research Centre, The University of Queensland, St Lucia, QLD 4072, Australia; s.phinn@uq.edu.au (S.P.); p.scarth@uq.edu.au (P.S.)
[2] Applied Agricultural Remote Sensing Centre, The University of New England, Armidale, NSW 2350, Australia
* Correspondence: andrew.clark@une.edu.au

Abstract: Data pre-processing for developing a generalised land use and land cover (LULC) deep learning model using earth observation data is important for the classification of a different date and/or sensor. However, it is unclear how to approach deep learning segmentation problems in earth observation data. In this paper, we trialled different methods of data preparation for Convolutional Neural Network (CNN) training and semantic segmentation of LULC features within aerial photography over the Wet Tropics and Atherton Tablelands, Queensland, Australia. This was conducted by trialling and ranking various training patch selection sampling strategies, patch and batch sizes, data augmentations and scaling and inference strategies. Our results showed: a stratified random sampling approach for producing training patches counteracted class imbalances; a smaller number of larger patches (small batch size) improves model accuracy; data augmentations and scaling are imperative in creating a generalised model able to accurately classify LULC features in imagery from a different date and sensor; and producing the output classification by averaging multiple grids of patches and three rotated versions of each patch produced a more accurate and aesthetic result. Combining the findings from the trials, we fully trained five models on the 2018 training image and applied the model to the 2015 test image. The output LULC classifications achieved an average kappa of 0.84, user accuracy of 0.81, and producer accuracy of 0.87. Future research using CNNs and earth observation data should implement the findings of this project to increase LULC model accuracy and transferability.

Keywords: convolutional neural network; deep learning; semantic segmentation; land use; land cover; aerial imagery

Citation: Clark, A.; Phinn, S.; Scarth, P. Pre-Processing Training Data Improves Accuracy and Generalisability of Convolutional Neural Network Based Landscape Semantic Segmentation. *Land* **2023**, *12*, 1268. https://doi.org/10.3390/land12071268

Academic Editors: Yimin Chen, Guohua Hu, Yujia Zhang, Xuecao Li and Brian Alan Johnson

Received: 28 April 2023
Revised: 12 June 2023
Accepted: 17 June 2023
Published: 21 June 2023

1. Introduction

1.1. Land Use and Land Cover Mapping

Remote sensing as a young science has already undergone several paradigm shifts, for example, from plain pixel-based analysis to subpixel analysis and geographic object-based image analysis [1]. Lately, the term 'big data', labelled as the fourth paradigm in science [2], has been used to describe challenges associated with data-intensive sciences. The steadily increasing volume of data, such as remotely sensed multispectral imagery, leads to general big data problems where methods are required to process and analyse the input data for efficient, generalised, transferrable and accurate information extraction [3,4].

Advances in earth observation technologies have provided efficient and cost-effective land use and land cover (LULC) mapping, encompassing a larger area with higher accuracy compared to traditional field surveys [5]. As a result, many programs around the world were established based on manual digitisation in a Geographic Information System (GIS) environment using image interpretation with assistance from other ancillary data [6].

Currently, there are operational LULC mapping programs in several countries. These include the Global Land Cover Characteristics Database from the United States Geological Survey, Co-ORdinated INformation on the Environment (CORINE) from the European Environmental Agency, GeoBase from the Canadian Council on Geomatics and Natural Resources and the Australian Land Use and Management Program. These programs involve extensive manual interpretation of imagery and other ancillary data to derive LULC, although there has been some automation of land cover classes.

1.2. Automated Land Use and Land Cover Classifications

Image classification has been fundamental in LULC analysis since the early days of the remote sensing discipline [7]. There have been many studies exploring classification techniques for LULC classification; however, it is still unclear which are the best classifiers [8].

Traditional approaches, such as maximum likelihood, fuzzy logic and object-oriented classifications, are referred to as shallow learning. These methods extract data based on spatial, spectral, textural, morphological and other cues [9]. However, shallow learning analytical techniques for extracting LULC information using high spatial resolution imagery but low spectral and temporal resolution cannot successfully separate land use classes due to similar spectral signatures between features [8].

In contrast, because deep learning is multi-layered and learns from the data itself, results can be significantly more accurate than those of shallow learning [10,11] and it has been shown to outperform manual human editing [12].

1.3. Deep Learning

Deep learning for image segmentation has recently become the superior classification technique for earth observation data, including the identification of LULC in very high-resolution (<1 m) data. This popularity is evident from the number of review papers attempting to bring order and clarity to the plethora of recent studies [13–16].

Convolutional Neural Networks (CNNs), a form of deep-learning, can utilise contextual information as well as spectral information to undertake image analysis. CNNs are the current network architecture of choice, with U-Net [17] being one of the most popular [13]. The U-Net has been used in multiple LULC studies, including forest coverage [18], urban studies such as building, road and car identification [19], and agricultural applications [20].

Numerous studies have shown deep learning techniques can successfully classify land use features; however, there are limited real-world mapping applications as most of the literature surveyed was constrained geographically or restricted to a standard set of training images, such as the University of California Merced Land Use Dataset. Furthermore, there are few studies that focus on generalisation and assess accuracy when applying the model to data from another time, sensor, or geographical area [16].

Challenges for Deep Learning Applications and Project Aims

There are several challenges that new projects must overcome to use deep learning techniques for automated data extraction from earth observations. Obtaining and processing training data is a major hurdle, although there is an abundance of imagery [21]. There are freely available existing datasets that can form the basis for training data, such as those discussed in Section 1.1. However, these products are produced at a continental scale with resolutions greater than 30 m and do not match the sub-metre spatial resolution imagery generally used with CNNs. Although some effort has been made recently to release higher resolution datasets (e.g., https://github.com/Seyed-Ali-Ahmadi/Awesome_Satellite_Benchmark_Datasets accessed on 16 June 2023).

A recent review article by [22] concluded that computer vision methods are yet to be unified and integrated with traditional earth observation analysis for LULC mapping. The article recommended that novel tools and approaches be developed that combine

computer vision technology with earth observation data for LULC mapping. In addition, Vali et al. [23] discusses the technical issues associated with applying deep learning to earth observation data, including data preparation.

In earth observation data and classifications, the disparity within and between classes causes class imbalances. Within a particular class, there can be a variety of factors, such as environmental and climatic influences, solar angle and clouds, and the influence of recent or absent rainfall, that need to be considered. Further to this, there can be variations within the class simply because of the elements from which it is composed. For example, the tree fruit class within the Australian Land Use and Management (ALUM) classification represents several different tree fruit crop types, such as papaya, banana and mango, which all have different leaf and growth structures. Although increasing the classification resolution to the commodity level assists with class consistency, this can present other challenges as a subtle variation between classes makes separation difficult using earth observation without additional information such as ground-based observations. A balance is needed between what training data can be collected from satellite or aerial imagery and creating a robust model to account for variation within and between classes.

Another challenging factor in remote sensing applications is the class representation of the landscape. Most applications tend to have one dominant class and several classes that make up only a small proportion of the landscape. Systematic or random generation of training patches (or image chips) will have very few training samples for underrepresented classes, resulting in poor classification for these. With one class dominating the image, it is highly likely that the classification algorithm will misclassify smaller classes without incurring a huge penalty. In addition, different areas within the training data class features result in features with small areas being less sampled than larger features, resulting in their poor classification.

To address LULC mapping in a consistent and repeatable way, this project aimed to establish standard training data processing recommendations that can be generally applied to high-resolution RGB earth observation data prior to training a deep learning model, including:

- how to sample the data;
- which patch size was the most effective;
- what effect the size of the batch of training data had on model training;
- how to ensure model transferability through data augmentations and scaling;
- how to create a more accurate and aesthetic classification by averaging the results of multiple prediction passes and augmentations.

Determining a standard set of pre-processing parameters for training data will assist future projects on how to approach deep learning segmentation problems.

2. Methodology

2.1. Project Area

The project area, located in North Queensland, Australia, encompasses the towns of Mareeba in the north, Atherton in the south and Dimbulah in the west (Figure 1).

The eastern part of the project area contains part of the world heritage-listed Wet Tropics rainforests, with the dominant land uses of production from relatively natural environments, conservation and natural environments, and production from irrigated agriculture and plantations (Table 1, Figure 1) [24].

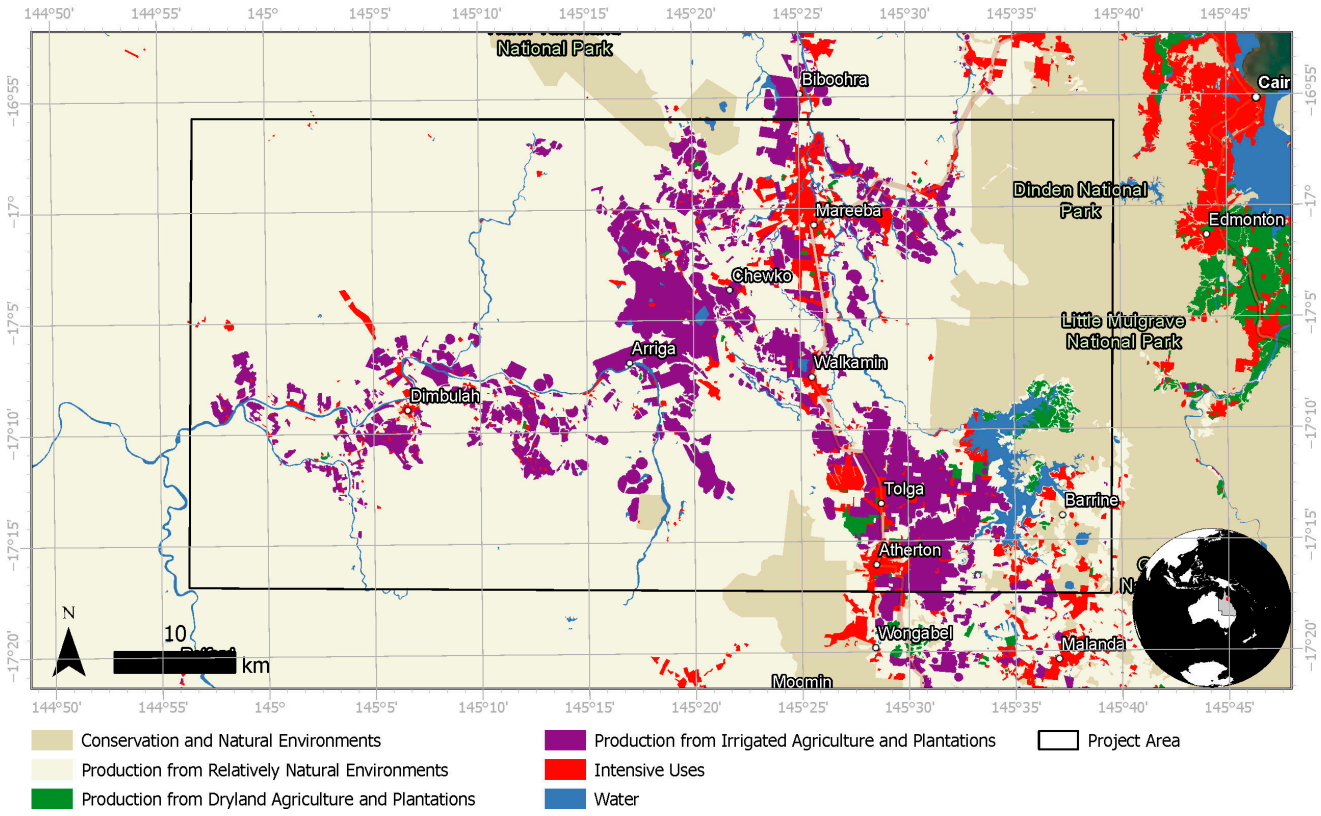

Figure 1. Land uses [24] and project area, North Queensland, Australia.

Table 1. Project area primary land uses [24].

Primary Land Use	Hectares	Proportion
Production from relatively natural environments	201,625	67.3%
Conservation and natural environments	40,404	13.5%
Production from irrigated agriculture and plantations	37,939	12.7%
Intensive uses	10,133	3.4%
Water	7219	2.4%
Production from dryland agriculture and plantations	2143	0.7%
Total	299,463	100.0%

2.2. Image Data

Two orthorectified aerial imagery mosaics acquired under the Queensland Government Spatial Imagery Subscription Plan were used in the project (Figure 2). The mosaic used for training data and assessing each trial was acquired between 1 and 27 August 2018, referred to here as the 2018 training image. The training image was mostly captured using a Vexcel Ultracam Eagle camera except for the southwestern corner, which was captured by an A3 Edge camera. The two cameras have different spectral properties, which can be seen in Figure 2b. The test mosaic was acquired between 17 July and 14 October 2015, using an A3 Edge camera, referred to here as the 2015 test image.

The images were captured with a mounted fixed-wing, three-band (true-colour) camera at spatial resolutions of 25 cm and 20 cm for 2015 and 2018, respectively. The data were provided as orthorectified mosaics. As shown in Figure 2, the quality of the imagery is not consistent across the project area. Unfortunately, the specific post-processing details were not listed within the supplied metadata; however, this was the highest resolution data available for the project area within the Queensland Government archive.

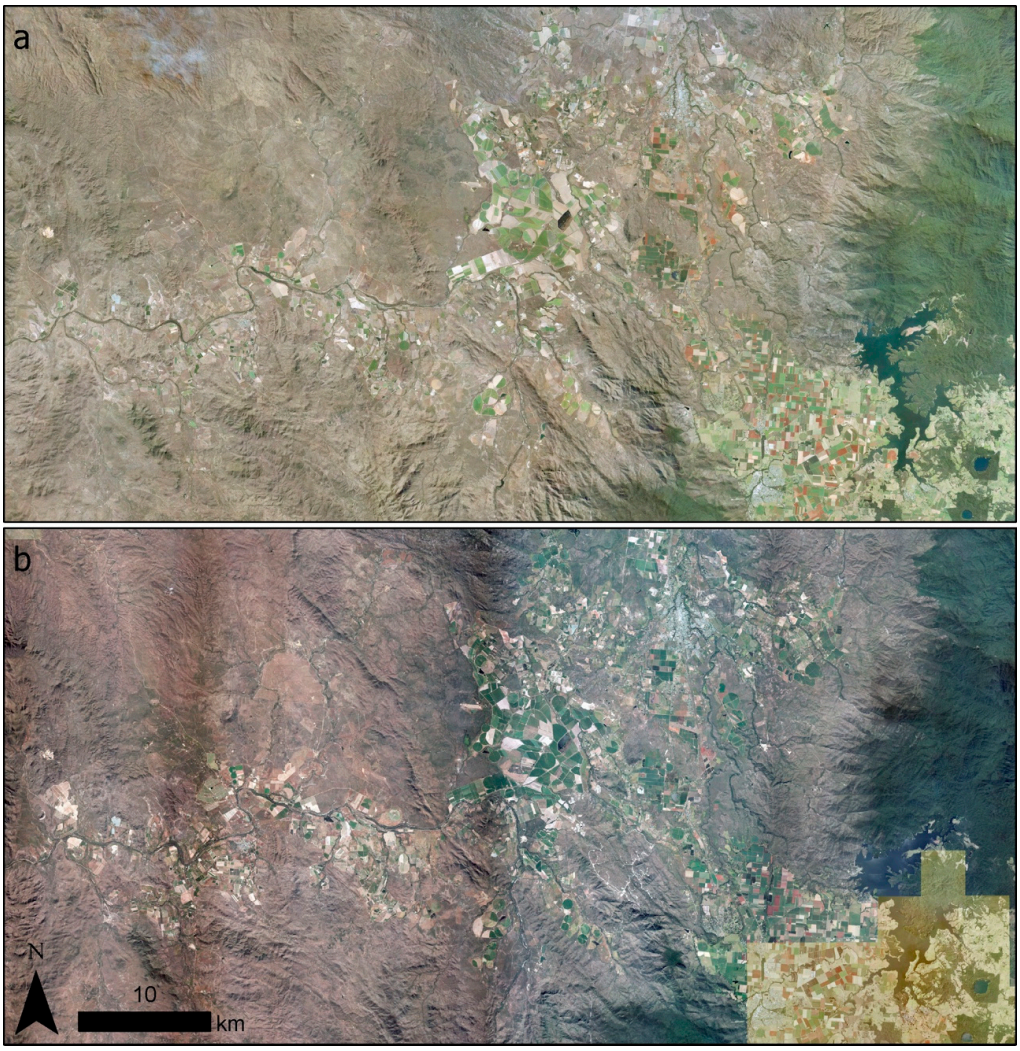

Figure 2. Project area orthorectified aerial imagery for 2015 (**a**) and 2018 (**b**). Data supplied by the Queensland Government.

The data were resampled to 50 cm using cubic convolution to reduce data volume, reduce overall training time, and ensure the resolution was consistent between images.

2.3. Data Collection

Eight LULC classes were selected, along with a ninth 'Other' class covering all other land uses. The chosen classes fall within the 'Production from irrigated agriculture and plantations' primary land use class presented in Table 1. These classes tend to be more dynamic compared to other classes such as 'Conservation and natural environments' or 'intensive uses', which are less likely to change.

Using ArcGIS Pro, the data were manually collected within the 2018 training image and 2015 test image by hand digitising polygons to best represent each feature's extent (e.g., the crop boundary). The classes included banana plantations, berry crops, forestry plantations, sugarcane crops, mature tree crops, young tree crops, tea tree plantations and vineyards, with all other remaining areas within the test area classified as other land uses. These classes were selected based on existing training data from previous work [20], prior knowledge of the area, ease of identification of the features within imagery, and the ability to digitise on-screen to a high level of detail.

The 2018 data was used for training, while the 2015 data was used as a comparison between the model and human classifications with the independent accuracy assessment detailed in Section 2.8.

2.4. Field Verification

A field trip to the project area was conducted in January 2020, where 1582 point-based roadside observations were made. These data were used in the verification and refinement of the training data. However, due to the time difference between the field observation and the imagery acquisition date (17 months), these data were not used to assess the accuracy of the final classification.

The observations were collected using an Apple iPad Pro running Collector for ArcGIS. The data were contained in a feature service and stored within the ArcGIS Online servers. For each point observation, the LULC type (e.g., banana crop) and growth stage (e.g., mature crop) were recorded. Optionally, additional fields allowed recording information such as land management (e.g., irrigation), observation photos, and other information within a comment field.

2.5. Deep Learning Trials

The objective of this project was to determine a standard set of training processes that can be generally applied to earth observation data. These processes include training data sampling strategies, patch size and how to feed these data to the GPU (batch size, data scaling and augmentations) for learning (Figure 3). Additionally, we trialled the derivation of the classifications based on the average of multiple inferences from the same test image. This was achieved through augmentation (image rotations) and by offsetting the start of the inference by half the number of pixels contained within a single patch.

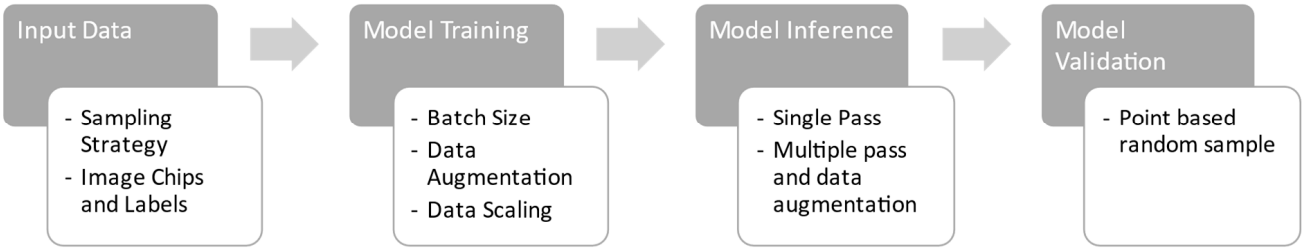

Figure 3. Project trials, stages and validation steps.

For this part of the project, all parameter trials were tested independently. Table 2 summarises the parameter trials, range of test values and default values when the parameter was not being tested. The only parameters altered within the trials were the ones to be tested. All other parameters remained consistent throughout the training. There was no default value for batch size, as this was optimised according to and consistent within each trial parameter.

Table 2. Parameter trials, number of patches, test values and defaults for the deep learning trials.

Parameter	Number of Training Patches	Test Values	Default
Batch size	3986	10, 50, 100, 150, 200, 250 and 280	*
Patch size and sampling strategy	2408–154,514	128^2, 256^2, 512^2 and 1024^2	512^2
Data Augmentations	22,830	True, False	False
Data Scaling	22,830	True, False	False
Multiple-Pass Prediction	22,830	True, False	False

* batch size was optimised for each trial.

Within each trial, multiple values of the parameters were tested by training five models for twenty epochs (or iterations of the training data) for each value. It was deemed that repeating the training a total of five times was sufficient to capture any random variance within the training process. Training the models for twenty epochs was not enough iterations of the data to produce a fully trained model; however, it was enough to give an

indication of training performance and examine the training progress consistently over all tests while reducing the resource load.

The results can only be compared against each other within the same parameter trial and are not an indication of the accuracy of the classification from a fully trained model. Training each trial five times for twenty epochs and assessing the accuracy gave an assessment of how well a particular set of parameters performed against the 2018 training image; however, we refrained from assessing the accuracy against the 2015 test image until we had fully trained models based on the optimised parameters. The method to fully train a model will be discussed in Section 2.10.

2.5.1. Batch Size

The batch size refers to the number of patches and labels that are processed on the GPU at one time and is limited by the size of the patches and GPU memory. Using a patch size of 512 × 512 pixels with three bands and corresponding nine-band labelled data, the maximum number of patches that could be processed at one time was limited to 280, as exceeding this value would cause an out-of-memory error. The trials for batch size consisted of batches ranging from 10 to 280 (Table 2). As this trial involved training 35 models, the number of patches was restricted to 3986 to limit time and resource requirements.

2.5.2. Patch Size and Sampling Strategy

Two training data sampling approaches were trialled: a systematic grid sampling strategy and a stratified random sampling approach based on area:

Grid sampling strategy

The grid sampling strategy is a systematic sampling approach. The image is divided into an even grid, with each grid cell being 1024 × 1024 pixels (512 × 512 m) in size. To reduce the number of patches, only grid cells that intersected the eight main training classes were retained. Patches that only intersected the 'other' class were excluded—data for this class was still contained within the target patches as 'other' land uses surrounded the eight main classes within the landscape.

To test for optimum patch size, these grid cells formed the basis for subsequent grid cell sizes. Each 1024 × 1024 pixel cell was divided into four cells to produce grid cell sizes of 512 × 512 pixels (256 m × 256 m). This process was repeated to produce a minimum patch size of 128 × 128 pixels (64 m × 64 m). Figure 4 shows the distribution of the training data.

This approach ensured the same data were fed to the CNN but split into different-sized patches. With every reduction in patch size, the number of patches increased fourfold.

Stratified random sampling strategy

The systematic grid sampling strategy is a comprehensive sampling strategy; however, it does not account for the imbalance in class areas with the training data dominated by larger classes. To account for this disproportion as well as ensure all features within a particular class are sampled at least once, a stratified random sampling approach based on the area was developed.

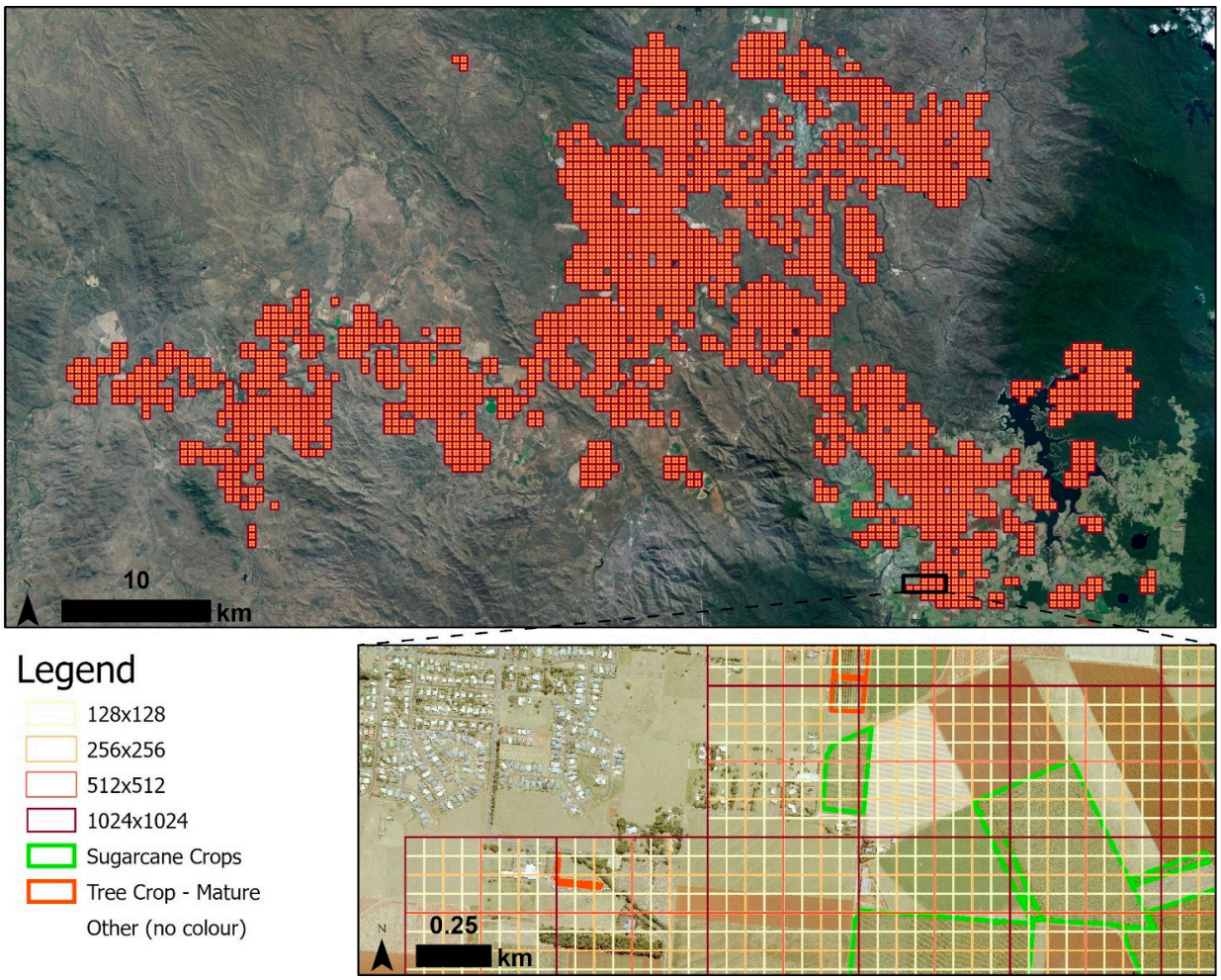

Figure 4. Patch layout from the grid-based sampling strategy. The training classes shown are derived from the manually derived training dataset.

For this sampling strategy, the number of patches for a particular class was calculated by multiplying the required number of patches (based on the result of the grid sampling strategy) by the log area of the target class and dividing by the sum of the log areas for all classes. The result was rounded up to the nearest integer (Equation (1)).

$$N_{cp} = \left\lceil N_p \frac{\log(a_c)}{\sum_n^i \log(a_c)} \right\rceil \quad (1)$$

where N_{cp} is the number of class patches, N_p is the total number of training patches, and a_c is the class area.

Each feature within the training data was sampled at least once to ensure all variations of the classes were captured. The number of patches generated within the feature was calculated by multiplying the number of class patches by the proportion of area that the feature represented of the total class, rounded up to the nearest integer (Equation (2)).

$$N_{fp} = \left\lceil N_{cp} \frac{a_f}{a_c} \right\rceil \quad (2)$$

where N_{fp} is the number of feature patches, N_{cp} is the number of class patches, a_f is the feature area, and a_c is the class area.

To generate the patch extent, a coordinate was randomly selected within the feature's geometry. The generated point formed the centroid of the patch.

To be consistent with the grid approach, a similar number of training patches (N_p) was produced. The stratified random sample approach rounded up the number of classes and feature patches and sampled every feature within the training data. As a result, the exact number of patches could not be matched without randomly excluding data, which may disproportionately affect the classes. Table 3 lists the difference in patch number for each patch size.

Table 3. Number of patches for patch size and sampling strategy trials.

Patch Size (Pixels)	Number of Patches		Difference	Difference (%)
	Stratified Method	Grid Method		
128 × 128	154,514	154,112	402	0.26%
256 × 256	38,533	38,528	5	0.01%
512 × 512	9635	9632	3	0.03%
1024 × 1024	2412	2408	4	0.17%

Table 4 shows the number of patches for each class and patch size. The table highlights the differences between the methods and shows that the number of patches in the larger classes (e.g., other and sugarcane crops) has decreased while the smaller classes (e.g., berry crops and vineyards) have increased.

Table 4. Number of patches per class for each patch size and sampling strategy trial (grid-based sampling and stratified random sampling).

Class	128 × 128 Pixels		256 × 256 Pixels		512 × 512 Pixels		1024 × 1024 Pixels	
	Grid	Stratified	Grid	Stratified	Grid	Stratified	Grid	Stratified
Banana Plantations	6155	17,509	1861	4535	637	1214	249	369
Berry Crops	387	14,050	129	3567	49	892	21	230
Other	132,194	98,019	35,955	3567	9519	892	2408	2410
Plantation Forestry	3604	16,565	1100	4191	360	1050	133	266
Sugarcane Crops	23,705	19,585	6994	5162	2247	1503	795	473
Tea Tree	766	14,795	274	3761	110	964	50	247
Tree Crops—Mature	26,360	24,152	9005	7151	3414	2531	1418	941
Tree Crops—Young	4361	17,304	1634	4636	722	1423	367	451
Vineyards	546	14,539	194	3693	84	933	43	241
Total	154,112	154,514	38,528	38,533	9632	9635	2408	2412

Figure 5 shows the spatial distribution of the stratified sampling method and illustrates the clustering of patches around classes with a smaller area.

When training a neural network, the model weights were updated after each batch of training data was fed to the GPU. Smaller batch sizes will update the model weights more often than larger batch sizes. However, larger batch sizes have more data to inform the update of the model weights. To control the effect of the batch size on the results, the number of iterations per epoch remained consistent for each of the patch size tests (Table 5). This was deduced by determining an optimal batch size for the patch size of 1024 × 1024 pixels, which was capable of fitting within the GPU memory (batch size of 16) and multiplying the result by 4, 4^2 and 4^3 to calculate the batch size value for 512 × 512 pixels, 256 × 256 pixels and 128 × 128 pixels, respectively. This resulted in 150.5 batches of data for each epoch (rounded to 151).

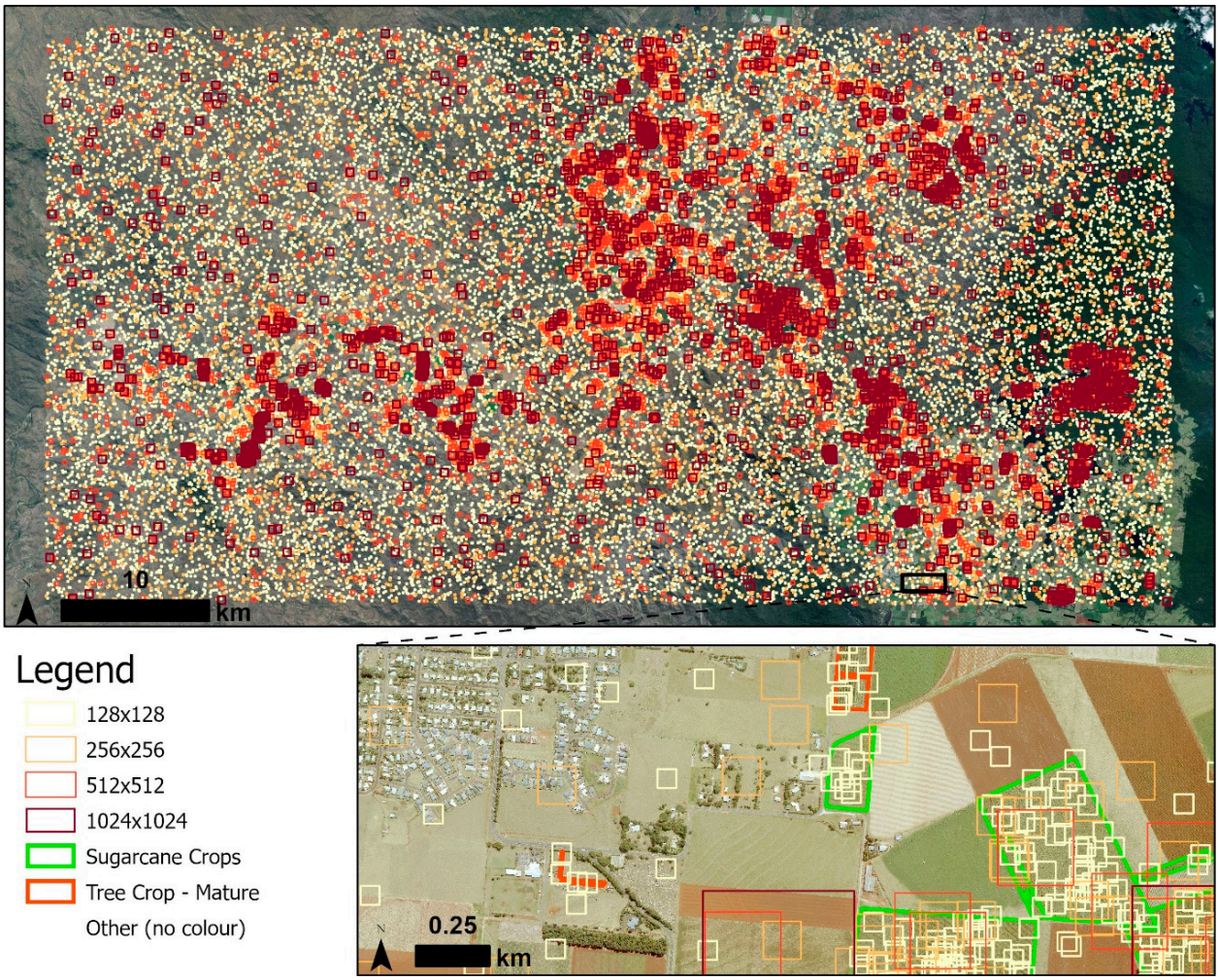

Figure 5. Patch layout from the stratified random sampling strategy. The training classes shown are derived from the manually derived training dataset.

Table 5. Parameters for setting up the log number patch experiments. The number of iterations is calculated by dividing the number of patches by the batch size and rounding to the nearest whole number.

Size (Pixels)	Number of Patches	Batch Size	Batches Per Epoch
1024 × 1024	2408	16	151
512 × 512	9632	64	151
256 × 256	38,528	256	151
128 × 128	154,112	1024	151

2.5.3. Data Augmentations

Aerial imagery can have poor calibration and varying quality and resolution, particularly between capture dates, because the same vendor, aircraft, camera and camera condition may not be used. As a result, spectral reflectance and spatial distortions can affect the appearance of features within the data. In addition, varying climatic conditions can also affect the spectral reflectance of features. To attempt to capture these variations, the training data can be augmented by flipping, rotating and changing the brightness of the image [25,26], which creates a more robust model for these image types and prevents overfitting of the data [14].

The Python package imgaug v0.4.0 (accessed on 16 June 2023) was used to apply random augmentations to the training data. The types of augmentation chosen were

dependent on whether they were deemed useful for remote sensing applications. The augmentations selected were based on:

- altering the contrast and colourations (gamma, sigmoid, AllChannelsCLASHE, linear, multiply and allChannelsHistogramEqualization);
- adding noise to the image (salt and pepper, multiply element-wise, additive Gaussian, additive Poisson and multiply—different for each channel);
- and altering the geometry and scale of the image by zooming and stretching the image (affine, elastic transformation, vertical and horizontal flips);
- adding blur and artificial clouds/fog/smoke to mimic varying environmental and climatic conditions, different resolutions, capture angles and aircraft roll effects, which are not always fully corrected in the provided imagery.

For each training image patch, one augmentation was picked randomly for each of the contrast, noise and geometric distortions, and 50% of the time either blur or artificial clouds/fog are applied. The augmentations were applied to each patch at every epoch, resulting in different versions of the patches on every iteration.

Figure 6 shows examples of the randomly applied augmentations. Note the size of the image provided does not represent the actual patch size used in this project but instead are to demonstrate the augmentation output.

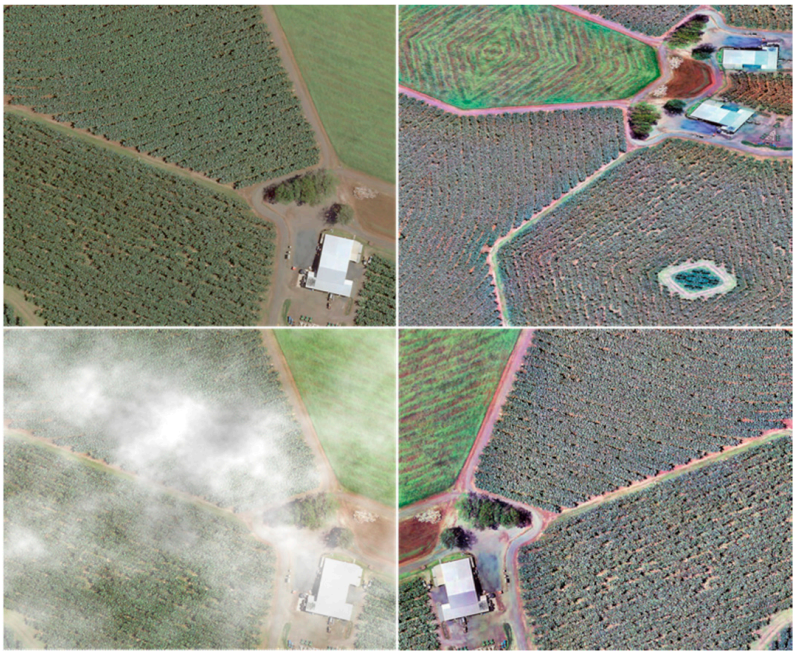

Figure 6. Example of data augmentations over an area consisting of banana plantations, including the original image (top left) and three augmentation versions. Note that these examples do not represent the patch size used in this project but are a demonstration of the augmentations used for each patch.

2.5.4. Data Scaling

Imagery from earth observation data can be supplied with a variety of pixel depths, including integers and floating point numbers in a variety of sizes, such as 8-bits or 32-bits, and can include negative or positive only values. Without patch data scaling, models trained with, for example, pixel values between −1 and 1, will not transfer to imagery with values between 0 and 255. The imagery used in this project was supplied with an 8-bit data type; however, the distribution of this data varied between the images (Figure 7). The data were scaled between 0 and 255 for each batch of patches.

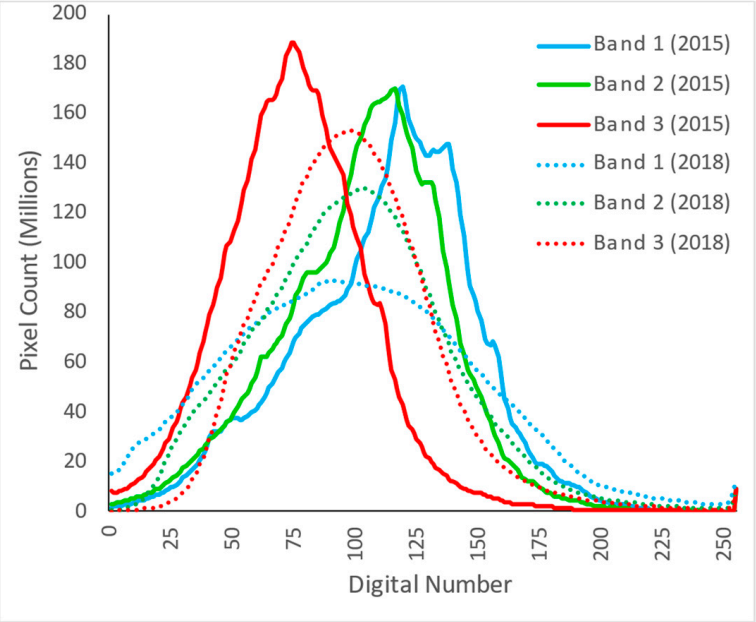

Figure 7. Pixel value distribution for 2015 (solid) and 2018 (dashes) image bands.

2.5.5. Multiple-Pass Prediction

It has been found in previous studies [27] that the edges of each image chip have lower accuracy than the centre region. To overcome this, a two-pass ensemble inference strategy was trialled. This was achieved by iteratively applying the model to the original image patch and averaging the resulting prediction from three rotated (augmented) versions. The second pass of predictions was offset by half a patch, resulting in the centre of the patches being located at the boundary of four of the first pass patches. The results from the two passes were combined using a weighted average based on distance, with pixels towards the centre of the patch given a higher weight than the pixels towards the edge.

2.5.6. Patch Image and Label Generation

For all trials except for patch size and sampling, 22,830 patch extents were generated spatially and stored within an ESRI shapefile. The extents were used to extract the training image patches. Corresponding labels were generated by converting the training polygon features to a raster representation covering the extent of the patches. The information for each class was added to a separate band within the label raster file known as one hot encoding, where 1 represents the presence of the class and 0 represents its absence. As a result, the label rasters consisted of nine bands representing each of the classes in alphabetical order.

2.6. Training

The purpose of the training stage was to allow the model to learn how to identify land use classes. This was achieved by iterating the training image patches and labels (training data) to determine their relevant colour, texture and context attributes [12]. Each trial consisted of 20 iterations (epochs) of the training data.

The aim of the training was to produce a model to label every pixel in the image through semantic segmentation using a CNN. The structure of the CNN was based on the U-Net architecture [17]. It consists of two parts: an encoding stage that down-samples the resolution of the input images and a decoding stage that up-samples and restores the images to their original resolution.

At each level of the encoding stages, two convolution operations were applied, and a 2×2 max pooling operation was used to down-sample the input images. The first level consisted of the original satellite image and label patches where a specified number of

filters were applied. For this project, we used 32 initial filters. At each subsequent level of the encoding side of the U-Net, the number of filters was doubled and the resolution halved until reaching the bottom level, where 512 filters were applied with a 16-fold reduction in spatial resolution (8 m) and pixels (e.g., 32 × 32 pixels for an original patch size of 512 × 512 pixels). Figure 8 is a graphical representation of the U-Net architecture from [20].

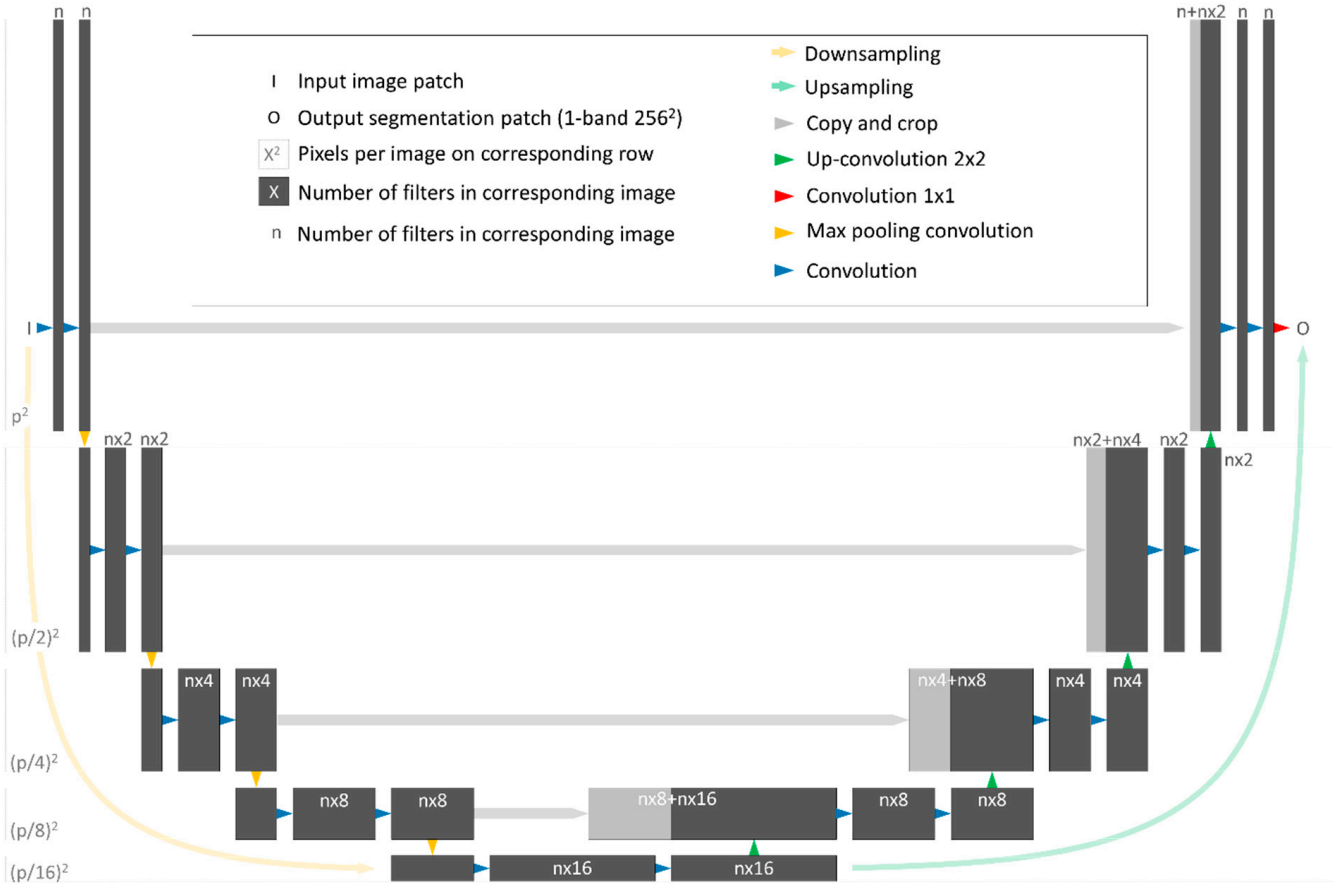

Figure 8. The U-Net architecture [17,20].

2.7. Prediction

Classifications for each trial were produced for the 2018 training image to assess how well a particular parameter learns from the training data. The classifications for the fully trained models based on the optimised parameters were produced for 2015 and 2018. The output from the model prediction is a raster with values between 0 and 1 for the nine classes represented in nine image bands. The prediction rasters are then flattened to a single-band thematic map, with the class containing the highest value considered the most probable feature for each pixel.

2.8. Accuracy Assessment

An independent accuracy assessment was conducted at the desktop by randomly generating 10,000 points in an unbiased sampling approach. At each point, an observation was made for the 2018 and 2015 imagery and classified according to the project classes. The data were stored within an ESRI Shapefile. Figure 9 shows the spatial distribution of data coloured according to the 2018 observation.

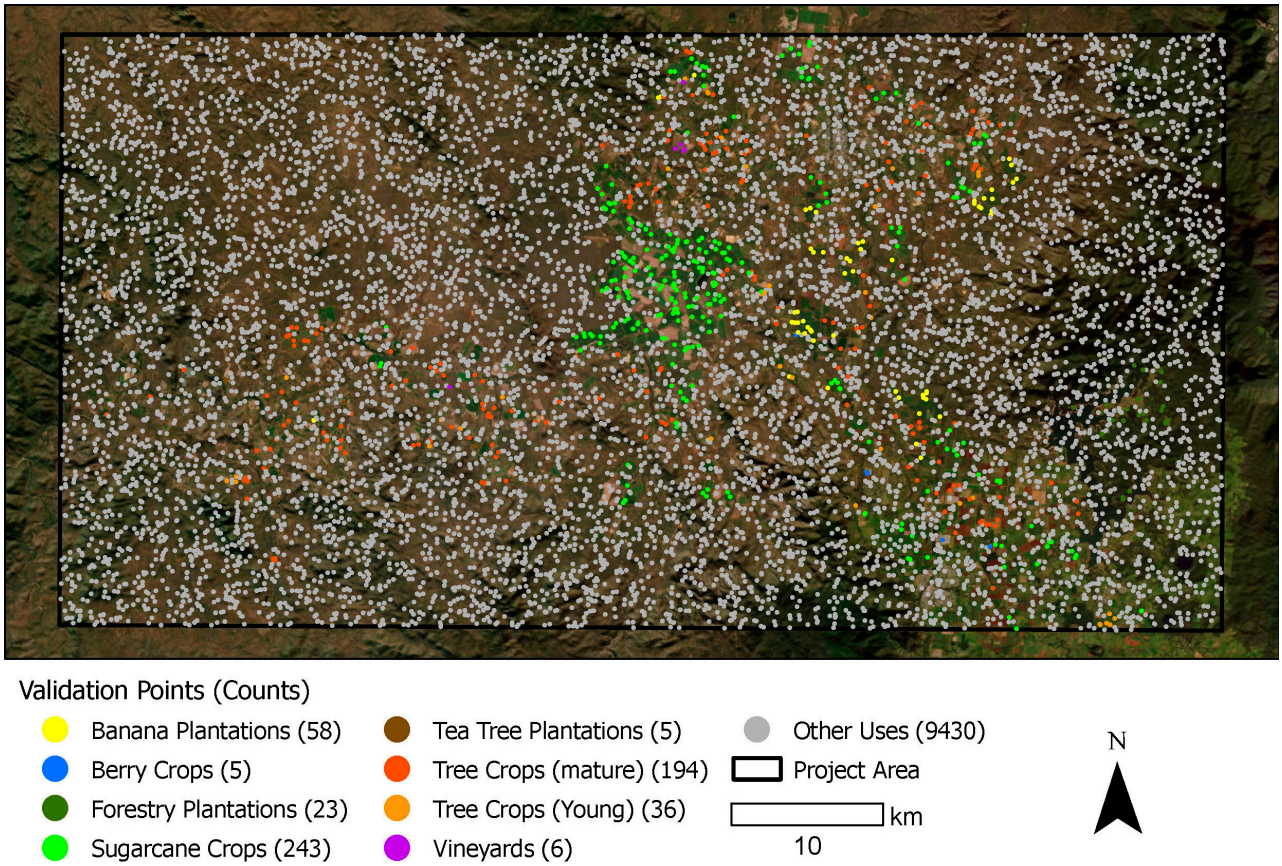

Figure 9. Validation points used to assess the accuracy of the model classification are coloured according to the 2018 observation value.

Although 10,000 points were generated, the disproportionate area between classes resulted in 94% of the points being located within the 'Other' class. This means smaller classes such as berry crops, vineyards and tea tree plantations only have five or six points. However, as the trials were repeated five times, this resulted in a minimum of 25-point observations used to calculate the statistics for each class. Although it would be ideal to assess the accuracy using additional points, time and resource considerations limited this capacity.

The hand-crafted training data and each trial classification were compared against the validation points, and accuracy was assessed by calculating the kappa and the user's accuracy (precision), producer's accuracy (recall) and F1-score metrics for each class.

2.9. Ranking the Trials

To assist in the interpretation of the results, the models were ranked by considering the reliability of the user's (recall) and producer's (precision) accuracy and model training time. To account for slight changes in computation time due to uncontrolled factors such as computing infrastructure load, the times were rounded up to the nearest 15 min. Time was only considered a factor if there was more than 15 min between the minimum and maximum.

For each test, the metrics were ranked from one to the total number of tests. A higher ranking represents higher accuracy or a lower computation time. The results were scored by adding the user's and producer's accuracies and twice the time ranks (resulting in the time having equal weighting as accuracy). Each test was assigned a final ranking based on this score.

2.10. Full Training of Top-Performing Models

Using the top trial rankings, full training was conducted to assess the highest possible result for the project area by training on 2018 data and applying the prediction to the 2015 data. As with the above trials, this was repeated five times but trained for 100 epochs.

2.11. Computing Infrastructure and Software

The Queensland Department of Environment and Science owns and operates High-Performance Computer (HPC) facilities. The HPC infrastructure consists of 2256 threads, 8.8TB of memory, eight Nvidia Tesla V100 GPUs and NVMe drives, which were used to process the training data, train the CNN model and create the model inference.

The processing of vector and image data used the Geospatial Data Abstraction Library (GDAL) version 3.1.0 (https://gdal.org/ accessed on 16 June 2023), and the deep learning part of the project utilised TensorFlow 2.1.0 [28]. Image augmentations used the Python library imgaug 0.4.0 (https://imgaug.readthedocs.io/en/latest/ accessed on 16 June 2023).

3. Results and Discussion

The aim of this project was to provide guidance on how to collect and process training data for use in deep learning projects involving earth observation data. The following sections present and discuss the results and provide recommendations on how other projects may best undertake the processing of the training data to produce the best possible results. Table 6 shows the results for all trials for the 2018 training image.

Table 6. Trial results for the 2018 training image.

Trial	Parameter	Average Training Time (h)	Kappa (95% CI)	Average F1—Score (95% CI)	Average User's Accuracy (Precision)	Average Producer's Accuracy (Recall)	Ranking
Human Derived	Manual	>200	0.96	0.95	0.97	0.92	-
Batch Size	10	1.2	0.67 (0.62–0.71)	0.68 (0.56–0.74)	0.63 (0.5–0.71)	0.84 (0.66–0.92)	5
	50	0.8	0.63 (0.62–0.65)	0.62 (0.6–0.64)	0.53 (0.51–0.55)	0.91 (0.89–0.92)	2
	100	0.8	0.55 (0.48–0.62)	0.56 (0.54–0.59)	0.49 (0.46–0.53)	0.85 (0.81–0.89)	1
	150	0.8	0.43 (0.31–0.49)	0.51 (0.44–0.56)	0.43 (0.38–0.49)	0.83 (0.81–0.86)	5
	200	0.8	0.41 (0.33–0.49)	0.45 (0.41–0.49)	0.38 (0.34–0.41)	0.82 (0.81–0.83)	4
	250	0.7	0.45 (0.33–0.51)	0.48 (0.43–0.51)	0.42 (0.37–0.44)	0.79 (0.77–0.8)	2
	280	0.7	0.36 (0.19–0.46)	0.41 (0.31–0.45)	0.36 (0.29–0.4)	0.75 (0.65–0.82)	5
Patch Size (Systematic)	128 × 128	3.7	0.69 (0.65–0.77)	0.49 (0.37–0.59)	0.53 (0.35–0.66)	0.51 (0.42–0.62)	2
	256 × 256	1.3	0.58 (0.4–0.72)	0.48 (0.34–0.56)	0.49 (0.34–0.61)	0.55 (0.42–0.63)	2
	512 × 512	1.2	0.7 (0.62–0.74)	0.49 (0.45–0.52)	0.53 (0.46–0.61)	0.53 (0.46–0.61)	1
	1024 × 1024	1.3	0.66 (0.48–0.75)	0.42 (0.41–0.45)	0.43 (0.4–0.46)	0.47 (0.42–0.52)	4
Patch Size (stratified-random)	128 × 128	3.7	0.42 (0.38–0.49)	0.46 (0.41–0.5)	0.38 (0.35–0.43)	0.86 (0.85–0.87)	4
	256 × 256	1.2	0.51 (0.47–0.58)	0.51 (0.47–0.57)	0.43 (0.4–0.49)	0.86 (0.85–0.86)	2
	512 × 512	1.2	0.65 (0.59–0.71)	0.62 (0.52–0.67)	0.59 (0.48–0.68)	0.77 (0.7–0.81)	1
	1024 × 1024	1.3	0.53 (0.31–0.66)	0.54 (0.38–0.62)	0.5 (0.36–0.56)	0.71 (0.52–0.82)	3
Data Augmentation	FALSE	2.8	0.73 (0.7–0.77)	0.7 (0.66–0.74)	0.64 (0.58–0.68)	0.87 (0.83–0.9)	1
	TRUE	8.9	0.49 (0.34–0.59)	0.5 (0.48–0.53)	0.43 (0.4–0.46)	0.75 (0.72–0.77)	2
Data Scaling	FALSE	2.1	0.69 (0.62–0.74)	0.64 (0.56–0.69)	0.59 (0.48–0.66)	0.78 (0.72–0.81)	2
	TRUE	2.1	0.77 (0.7–0.79)	0.74 (0.68–0.78)	0.68 (0.6–0.74)	0.88 (0.83–0.91)	1
Multiple-Pass Prediction	Single	8.9	0.49 (0.34–0.59)	0.5 (0.48–0.53)	0.43 (0.4–0.46)	0.75 (0.72–0.77)	2
	Multiple	8.9	0.55 (0.4–0.65)	0.52 (0.48–0.53)	0.45 (0.41–0.48)	0.75 (0.69–0.81)	1

3.1. Training Data

The collection of training data took over 200 h to hand digitise. This is a significant challenge for many semantic segmentation applications [29] and potentially negates any benefits of deep learning approaches. The availability of the existing banana dataset [20] assisted greatly, and it is recommended that future studies leverage existing datasets where possible.

Table 7 lists the 2018 training data collected for this project. Approximately 94% of the project area consisted of the 'other' class, while only 0.03% of the area contained berry crops. This class imbalance can be typical for projects classifying LULC using earth observation data.

Table 7. Number of features, area and proportion of the project area for each class in 2018.

Name	Feature Count	Area (ha)	Area (%)
Banana Plantation	243	1860	0.62
Berry Crops	69	92	0.03
Forestry Plantation	118	981	0.33
Sugarcane Crop	515	7621	2.54
Tea Tree Plantation	42	188	0.06
Tree Crop—Mature	2289	6249	2.09
Tree Crop—Young	280	988	0.33
Vineyards	33	146	0.05
Other	323	281,344	93.95
Total	3912	299,471	100.00

The ability of a model to successfully train with high accuracy is reliant on the accuracy of the training data. If the training data is of poor quality, the model may not be able to determine the ideal weights for the model neurons to achieve the highest accuracy for classification. To a certain extent, CNN models may account for some level of error in the training data but may result in the model being penalised for achieving higher accuracy than the data used to assess its performance [21].

The training data used in the project has its own inaccuracies, as maintaining focus while hand digitising features over extended periods of time can be problematic [30]. As shown in Tables 6 and 8, the human-derived training data achieved an F1-Score of 0.95. Analysing these data at the class level revealed some individual classes, such as the young tree crops class, which achieved an F1-Score of only 0.73 (Table 8), which limits the ability of the model to identify this class.

Table 8. Per-class accuracy of the human-derived classifications for 2018 and 2015.

Class	2018			2015		
	F1-Score	User's (Precision)	Producer's (Recall)	F1-Score	User's (Precision)	Producer's (Recall)
Banana Plantations	1.0000	1.0000	1.0000	0.9908	0.9818	1.0000
Berry Crops	1.0000	1.0000	1.0000	0.8571	1.0000	0.7500
Plantation Forestry	0.8400	0.9545	0.7500	0.8780	1.0000	0.7826
Sugarcane Crops	0.9839	1.0000	0.9683	0.9204	0.9946	0.8565
Tea Tree Plantation	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Tree Crop—Mature	0.9602	0.9361	0.9856	0.9403	0.9141	0.9679
Tree Crop—Young	0.7297	0.8710	0.6279	0.6897	0.8696	0.5714
Vineyards	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Other	0.9981	0.9973	0.9989	0.9968	0.9950	0.9987
Total	0.9458	0.9732	0.9256	0.9192	0.9728	0.8808
Kappa	0.9617			0.9293		

During the training data collection process, separating young and mature tree crops was extremely subjective and quite difficult, which was reflected in the accuracy assessment of the training data and, as a result, in the accuracy of the resulting models outlined below. As a comparison, the banana plantation class was based on a previous deep learning model developed by [20], resulting in an F1-Score of 1 for 2018 and 0.99 for 2015 (Table 8). The classes with smaller areas, such as tea tree plantations and vineyards, had high accuracy as these features were easily identifiable.

The sugarcane crop class was complicated by several challenges. Firstly, the 2018 training image was captured in August, early in the sugarcane harvest season, when the cane was either fully mature or yet to be planted. The dates for the 2015 test image were captured as late as October when some of the fields harvested early in the season contained young sugarcane. The collection of the training data for 2015 was consistent with the 2018 training data, and only mature sugarcane was included. However, for the accuracy assessment points, we decided to call the validation point sugarcane if the canopy of the crop was predominantly closed.

The sugarcane crop class was further complicated by the presence of maize crops in the southeast of the project area. Maize can look similar to sugarcane and was only separated during the training and validation data collection through local knowledge and identification of farm management practices only seen at a broad scale and not within a single image patch. Ideally, maize should be included as a separate class; however, due to time constraints, it remained part of the 'other' class. Additional training data may assist with the differentiation of these classes but was not conducted as part of this project. There were also areas of abandoned sugarcane crops that were not in production but were still identified by the models.

3.2. Batch Size

The batch size results indicate higher accuracy using smaller batch sizes (Table 6). During the training process, model weights are updated at the end of every batch, which results in models with smaller batch sizes updating their weight more often, likely resulting in more refinement of model weights and faster convergence. These results are consistent with [31], who recommended using lower learning rates with small batch sizes, although data scaling and augmentation may assist training when using larger batch sizes [32].

Based on the results of batch size in combination with patch size, it is recommended to increase patch size while decreasing batch size to achieve higher accuracy.

3.3. Patch Size and Sampling Strategy

A challenging factor within remote sensing applications is class imbalance from over or underrepresented classes in the training data [16]. For the project area in this project, the 'other' class represented 94% of the area, whereas the berry crop class only represented 0.03%. A systematic or random generation of training patches will have very few training samples for underrepresented classes, resulting in their poorer classification. In addition, different areas within class features mean features with smaller areas will be sampled less than larger features, resulting in their poorer classification.

Other studies have applied class weighting [33]; however, this does not solve the problem of under-sampled classes and features. Classes with smaller areas will still contain very few patches and remain underrepresented within the training data.

Results from the systematic grid sampling method (Table 6) did not indicate a significant increase in accuracy for each patch size. The stratified random sampling method showed the 512 × 512-pixel patch size produced higher accuracy for both the kappa statistic and F1-Score compared to other patch sizes. The kappa statistic indicated a slight reduction in accuracy compared to the grid sampling strategy (2018: −0.06); however, the F1-Score improved (2018: +0.13).

The stratified random sample approach is a workable solution to overcome the class imbalance issue. The overall results for the patch size and sampling strategies do not indicate a significant improvement in accuracy between the trials; however, the largest improvements were produced in the smaller classes, particularly for larger patch sizes. Figure 10 shows an example of this improvement for the vine class. The grid sampling strategy tests showed inferior performance, with some models not being able to classify the vine features. In contrast, the stratified random sample strategy was able to detect the vines in most cases with some degree of accuracy. Based on these results, the best performance was achieved with a patch size of 512 × 512 pixels.

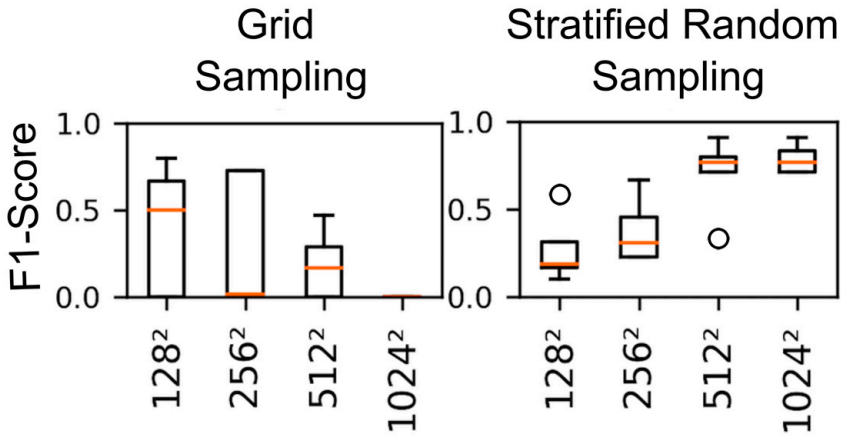

Figure 10. Box and whisker plots showing grid-based and stratified random sampling strategies for the vineyard class F1-scores. The box represents the first and third quartiles of the data, and the whiskers define the 1.5 inter-quartile range. The orange line represents the median, and the 'o' symbol represents outliers.

3.4. Data Augmentation

The data augmentation trials (Table 6) decreased model accuracy (kappa: −0.24; F1-Score: −0.2) and increased training time. This result is expected as the random augmentations have avoided overfitting the model to the training data by presenting an altered version of the training data each time the data is loaded. As a result, we decided to assess the model on the 2015 test image (Table 9). We found that when comparing models trained with and without augmentations, there was a >0.24 increase in kappa and a 0.19 increase in F1-Score. This demonstrates how to avoid overfitting the model to the training data to increase its transferability to unseen data.

Table 9. Augmentation trial results for the 2015 test image.

Parameter	Average Training Time (h)	Kappa (95% CI)	Average F1-Score (95% CI)	Average User's Accuracy (Precision)	Average Producer's Accuracy (Recall)	Ranking
False	2.8	0.12 (0.09–0.15)	0.21 (0.18–0.25)	0.22 (0.18–0.27)	0.35 (0.31–0.39)	2
True	8.9	0.36 (0.27–0.41)	0.4 (0.38–0.41)	0.37 (0.33–0.4)	0.61 (0.57–0.65)	1

Although applying augmentations increased the training time (threefold), it created a more robust model, which allowed for better transferability to other data.

The results from this trial indicate the importance of using a range of image augmentations to alter image perspective, colour and brightness. The pixel value variations within the 2018 training image and 2015 test images result from different camera configurations, post-processing of the image tiles into a seamless mosaic, and atmospheric and climatic conditions. These are all typical occurrences for earth observation data, particularly when using high-spatial-resolution data from aerial photography or satellites.

It is recommended that any project attempting to ensure model transferability not only to a different time but also to a different sensor or geographic region implement data augmentations.

3.5. Data Scaling

Table 6 shows the kappa statistic improved by 0.08 and the F1-Score by 0.1 when patch data scaling was applied. Although only one type of scaling was tested for this project, other scaling and normalising options are possible, such as scaling between zero and one. We used the range of 0–255 to maintain compatibility with the imgaug library,

which recommends pixel values in this range. We recommend the implementation of data scaling, particularly for projects using multiple imagery sensors.

3.6. Multiple-Pass Prediction

Comparing single and multiple-pass prediction methods (Table 6) shows the multiple-pass method marginally improves prediction accuracy (kappa difference: +0.06; F1-Score difference: +0.02). However, the multiple-pass classification is more aesthetic with the elimination of patch edge effects (Figure 11).

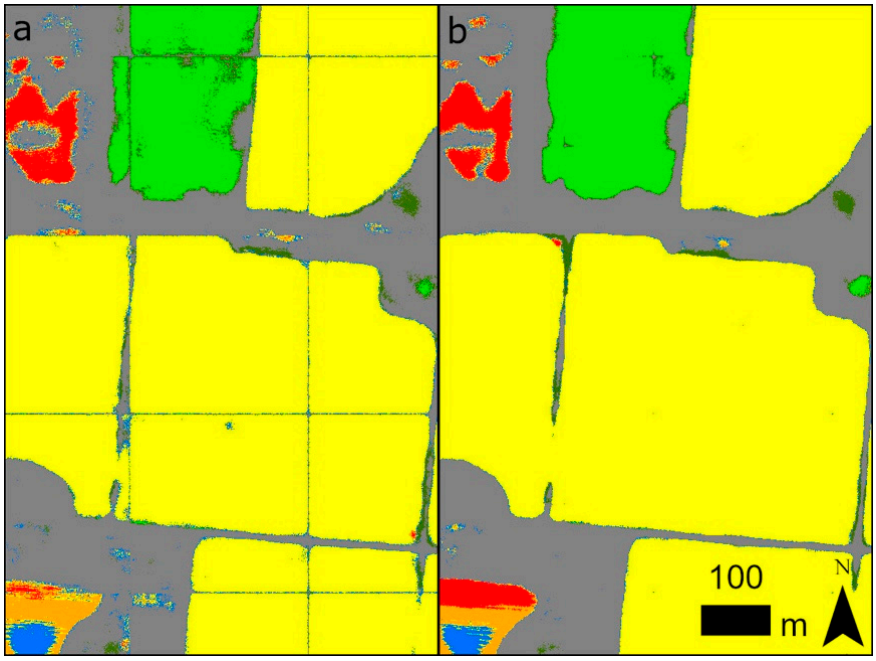

Figure 11. An example of a single pass prediction (**a**) compared to a multiple-pass prediction and augmentations (**b**).

Although only a marginal improvement in accuracy resulted from the multiple-pass method, this is recommended as it results in a smoother final classification.

3.7. Full Training

The objective of this project was to determine the optimal pre-processing steps to maximise the transferability of a model to a different sensor at a different date. Based on the results of the trials, five models were trained for 100 epochs using the parameters presented in Table 10.

Table 10. Parameters used for the 100-epoch training trials.

Parameter	Value
Patch size (pixels)	512 × 512
Sampling strategy	Stratified random sample (area)
Number of patches	22,830
Batch size	20
Data Augmentations	True
Data Scaling	True
Multiple-Pass Prediction	True

Training the models for 100 epochs resulted in the models achieving a kappa statistic of 0.9 (0.89–0.91) and 0.84 (0.82–0.87), a user accuracy of 0.8 (0.78–0.83) and 0.78 (0.76–0.8), and a producer accuracy of 0.98 (0.98–0.98) and 0.87 (0.85–0.9) for 2018 and 2015, respectively

(Table 11). As the models now contain the optimal pre-processing steps for the training data and are fully trained, we will no longer discuss the 2018 results.

Table 11. Resulting accuracy measures for the 100-epoch training trials.

Image	Kappa (95% CI)	Average F1-Score (95% CI)	Average User's (95% CI) (Precision)	Average Producer's (95% CI) (Recall)
2018	0.9 (0.89–0.91)	0.87 (0.86–0.89)	0.8 (0.78–0.83)	0.98 (0.98–0.98)
2015	0.84 (0.82–0.87)	0.81 (0.79–0.84)	0.78 (0.76–0.8)	0.87 (0.85–0.9)

Figure 12 shows the confusion matrix for 2015. The model performs well at finding all land use features; however, there is some confusion between the tree crop classes (mature and young) and between the other classes and the sugarcane, tree crops (mature and young), vineyards and tea tree classes.

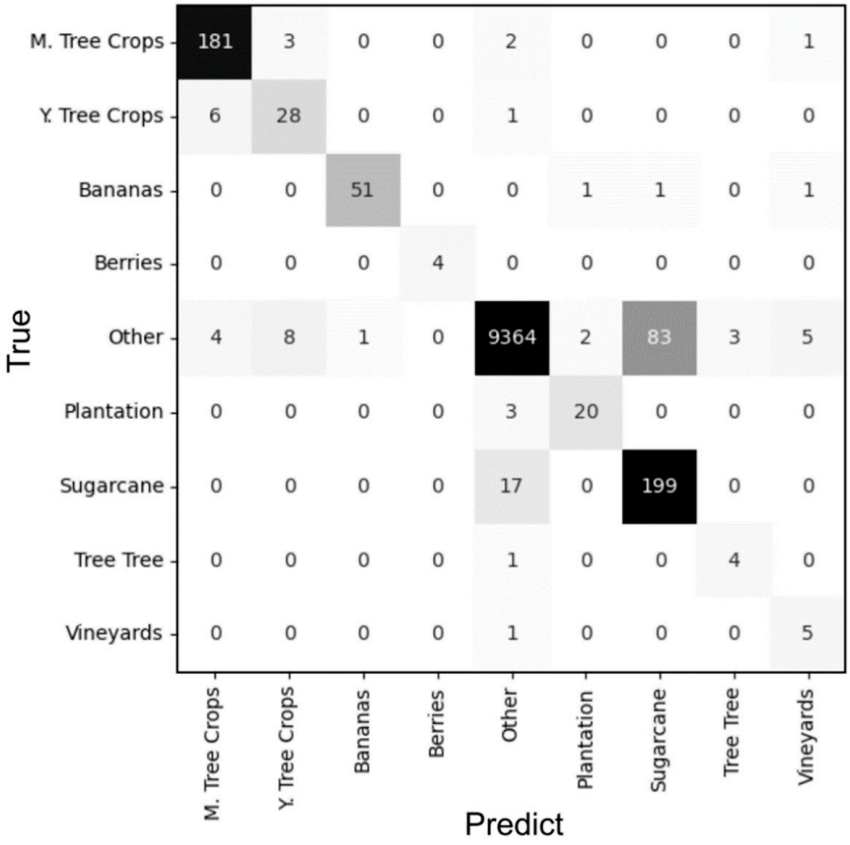

Figure 12. Full training trial number five, 2015 classification confusion matrix.

The per-class results for all five trials (Figure 13) showed most classes achieved an accuracy >70% except for the vineyards class. The main confusion was with areas of fallow. The 2018 training data contained many noticeably young vineyards where vines were barely evident in the image. This resulted in some vineyard features resembling areas of ploughed fallow. Only 0.5% of the project area contained the vineyard class (Table 7), which was the second-smallest class in the area. Berries were the smallest class; however, most berries in the project area were contained within a greenhouse, which makes these features easily identifiable within the imagery.

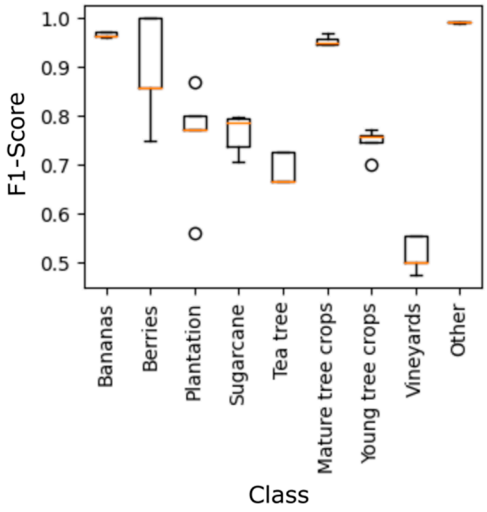

Figure 13. Box and whisker plots for F1-scores for each class from the five fully trained models. The box represents the first and third quartiles of the data, and the whiskers define the 1.5 inter-quartile range. The orange line represents the median, and the 'o' symbol represents outliers.

Figure 14 shows the human-derived (Figure 14a) and 2015 (Figure 14b) output classifications. At this scale, confusion between sugarcane crops and other land uses is evident.

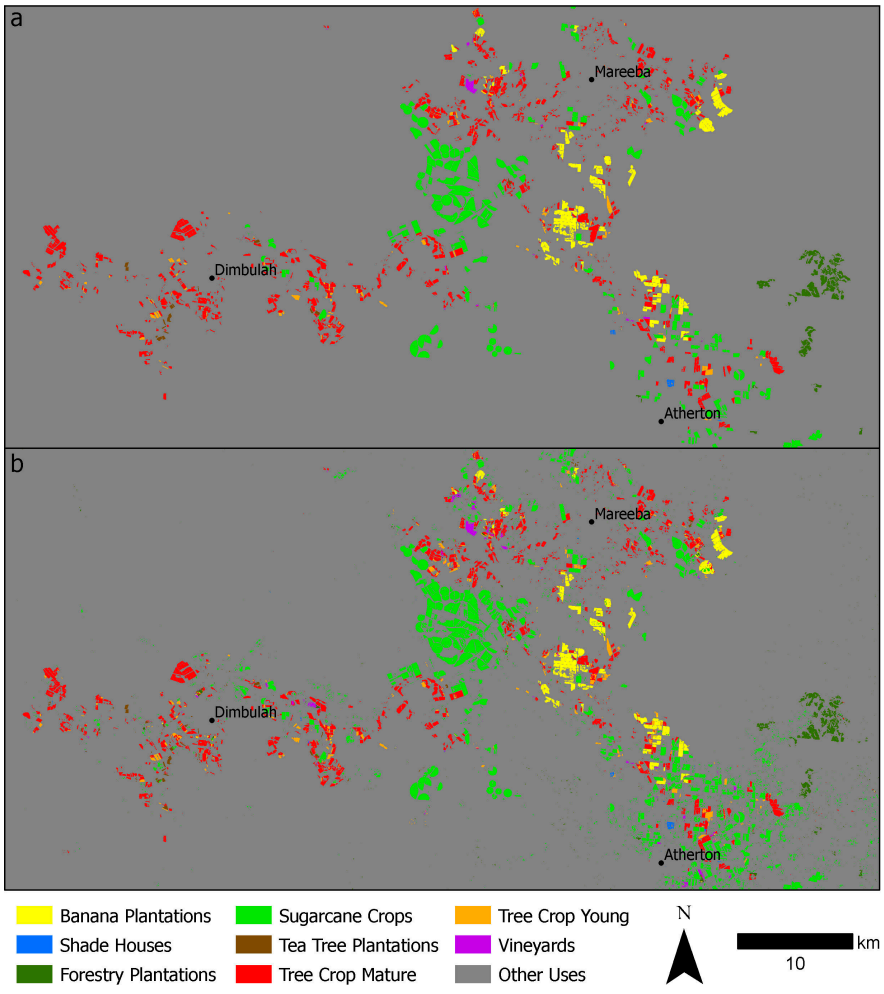

Figure 14. Comparison of the project area for human classification (**a**) and prediction results for 2015 (**b**).

Figure 15 shows the 2015 and 2018 imagery, hand-digitised 2015 validation and 2018 training data, and resulting classifications for the top performing of the five fully trained models. This model achieved a kappa of 0.87 and 0.84 and an F1-score of 0.84 and 0.77 for 2018 and 2015, respectively.

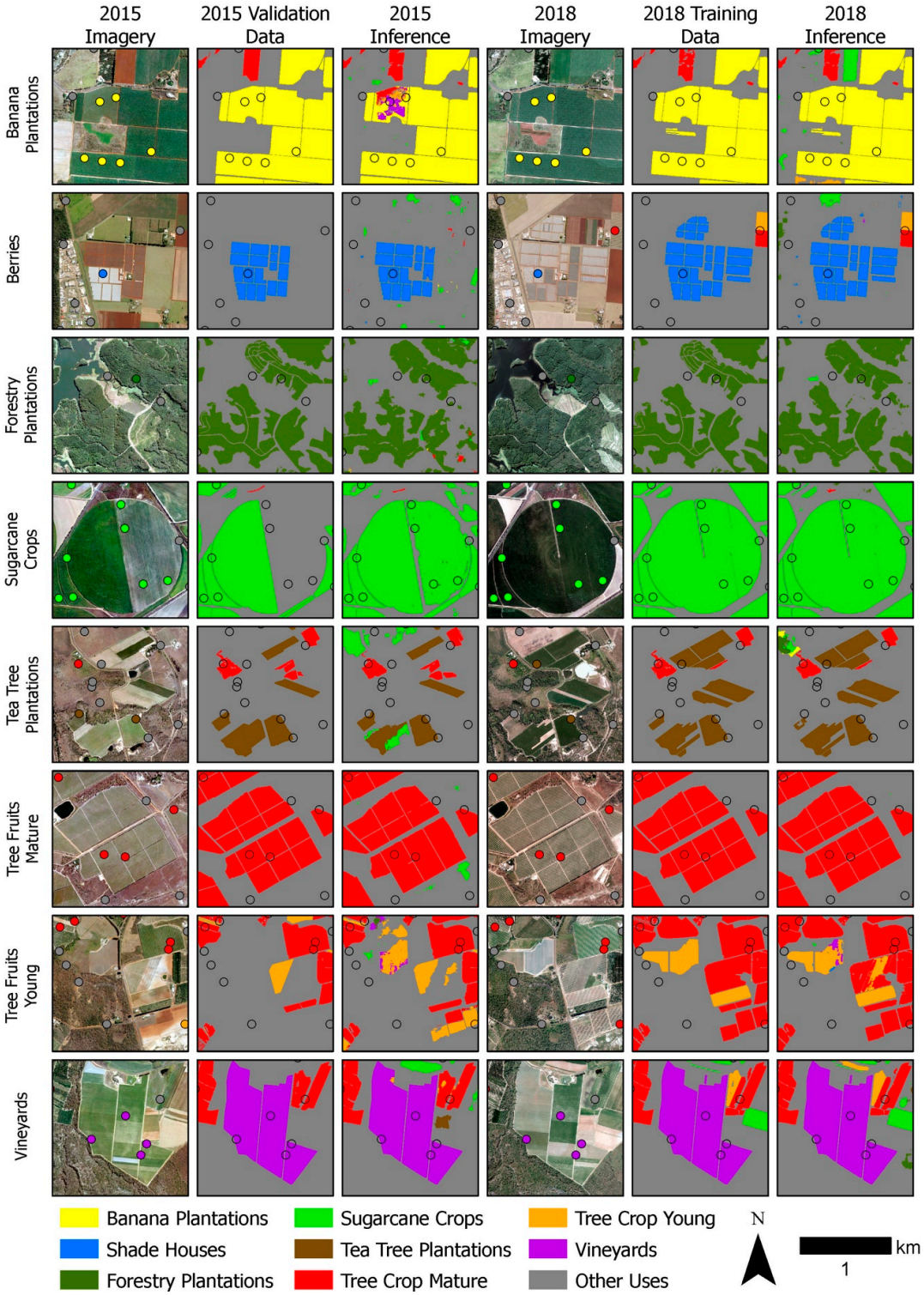

Figure 15. The 2015 and 2018 imagery, manual classification, and prediction results for a model trained for 100 epochs. The 2018 manual classification was used to derive the training data for the model. The circles indicate the location of the accuracy assessment points, with the colour in the imagery columns representing the class according to the legend.

The results showed some inaccuracies in the training and validation data. The 2015 manually classified data for the young tree crop, as shown in Figure 15, showed areas of missed tree crops that were identified by the model classification, consistent with the findings of other studies using noisy data [29]. The young tree crop example also identified an area of confusion between young and mature tree crops in the 2018 training data, which led to confusion with the model results (Figure 12).

Figure 15 also demonstrates an area of emerging sugarcane crops in 2015 that was not included in the manual classification as it was deemed the canopy had not fully closed as discussed in Section 3.1. In addition, it is also evident in Figure 15 that sugarcane crops are misclassified with the 'other' class, as demonstrated in Figure 12.

4. Limitations and Future Research

This project restricted the analysis to three-band, 50-centimetre imagery. These results may not be applicable for training a model for earth observation applications at different spatial and spectral resolutions. Convolutional neural networks are suited to higher-resolution data (<1 m) [34,35], although some success has been achieved using earth-i [18] and sentinel data [36]. It is expected that the recommendations presented in this project, such as sampling strategy, data augmentations, and multiple-pass prediction, will be applicable at different spatial and spectral resolutions; however, spatially specific aspects such as patch size (and as a result batch size) will need to be re-evaluated.

Additional research could investigate different approaches to producing the output classifications. In this project, we have used a multiple-pass strategy involving overlapping patches and a weighted mean to counteract patch edge effects when producing the output classification; however, there are alternatives such as trimming edge pixels and ensuring the patches overlap by the same number of pixels as presented in [18].

There are several outstanding questions that need to be addressed in future research. First, we did not classify all LULC features within the project area due to time restrictions. Future work should analyse the effect on model accuracy of additional LULC classes. It would also be interesting to compare the training time and model accuracy when a model is created for each class separately in contrast to having one model for all classes.

In this project, we compared generating the output classification using a single-pass strategy to a multiple-pass with augmentations. Although the multiple-pass method produced a more accurate and aesthetic classification, we did not compare the additional time it took to generate the classification. We did not examine if averaging the patch classification with three augmented versions has any advantage over one or two augmented versions. Producing outputs efficiently is imperative for the timely delivery of broad-scale LULC classifications.

The production of LULC classifications often relies on additional ancillary data and the existing experience of the skilled professional. The integration of additional ancillary information such as, for example, climate, elevation and soil information may assist in the prediction of LULC features, specifically over broad areas.

5. Conclusions

The objective of this project was to better understand how to prepare earth observation data for training a multi-class deep learning model to assist in the integration of traditional earth observation analysis for LULC mapping as recommended by [22]. Firstly, we recommend the use of existing datasets where available. The collection of training data for this project took hundreds of hours, even with existing datasets and prior knowledge of the area. Although freely available datasets usually do not match the spatial resolution used in projects implementing CNNs, the training of deep learning models can tolerate a certain level of noise within the data, and, in some cases, the trained model may have a higher level of accuracy compared to the original training data.

The most substantial improvements in the transferability of the model from the 2018 training image to the 2015 test image resulted from image augmentations and scaling of the

data. Data augmentation and scaling are imperative to avoid overfitting the model to the training data, model generalisation and transferability, and therefore are recommended.

Compared to the grid sampling approach, the stratified random sampling approach for generating image patches substantially increased inaccuracy for small classes in our imbalanced training dataset. This approach, although not improving overall model accuracy metrics, substantially improved accuracy in detecting classes that represent only a small proportion of the landscape. For projects with class imbalances, it is recommended that this sampling strategy be implemented.

When producing image patches, we recommend generating larger patch sizes and training with a lower batch number rather than smaller patches with larger batches.

Applying the model to imagery using different perspectives through patch rotations and applying for a second pass over the prediction image with a half patch offset improves the output accuracy and creates a more aesthetically pleasing classification.

Author Contributions: Formal analysis, A.C.; investigation, A.C.; data curation, A.C.; writing—original draft preparation, A.C.; writing—review and editing, A.C., S.P. and P.S.; visualization, A.C.; supervision, S.P. and P.S.; project administration, A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The aerial imagery data used to support the findings of this study were supplied by the Queensland Government under license and so cannot be made freely available. Requests for access to these data should be made to SIProductDelivery@resources.qld.gov.au. Code related to augmentations and the U-Net is located: https://github.com/clarka1/training_data_processing/tree/main (accessed on 16 June 2023).

Acknowledgments: The authors would like to acknowledge the support from the Australian Government Research Training Program Scholarship and the Queensland Government for supplying the data used in this paper and for the use of the High-Performance Computing Infrastructure.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Blaschke, T.; Hay, G.J.; Kelly, M.; Lang, S.; Hofmann, P.; Addink, E.; Queiroz Feitosa, R.; van der Meer, F.; van der Werff, H.; van Coillie, F.; et al. Geographic Object-Based Image Analysis—Towards a New Paradigm. *ISPRS J. Photogramm. Remote Sens.* **2014**, *87*, 180–191. [CrossRef] [PubMed]
2. Hey, T.; Tansley, S.; Tolle, K.; Gray, J. *The Fourth Paradigm: Data-Intensive Scientific Discovery*; Microsoft Research: Redomon, WA, USA, 2009; ISBN 978-0-9825442-0-4.
3. Cheng, G.; Han, J.; Guo, L.; Liu, Z.; Bu, S.; Ren, J. Effective and Efficient Midlevel Visual Elements-Oriented Land-Use Classification Using VHR Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 4238–4249. [CrossRef]
4. Ma, Y.; Wu, H.; Wang, L.; Huang, B.; Ranjan, R.; Zomaya, A.; Jie, W. Remote Sensing Big Data Computing: Challenges and Opportunities. *Future Gener. Comput. Syst.* **2015**, *51*, 47–60. [CrossRef]
5. Bai, X.; Sharma, R.C.; Tateishi, R.; Kondoh, A.; Wuliangha, B.; Tana, G. A Detailed and High-Resolution Land Use and Land Cover Change Analysis over the Past 16 Years in the Horqin Sandy Land, Inner Mongolia. *Math. Probl. Eng.* **2017**, *2017*, 1–13. [CrossRef]
6. Lillesand, T.M.; Kiefer, R.W.; Chipman, J.W. *Remote Sensing and Image Interpretation*, 7th ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2015; ISBN 978-1-118-34328-9.
7. Jensen, J.R. *Remote Sensing of the Environment: An Earth Resource Perspective*, 2nd ed.; Prentice Hall Series in Geographic Information Science; Pearson Prentice Hall: Upper Saddle River, NJ, USA, 2007; ISBN 978-0-13-188950-7.
8. Pandey, P.C.; Koutsias, N.; Petropoulos, G.P.; Srivastava, P.K.; Dor, E.B. Land Use/Land Cover in View of Earth Observation: Data Sources, Input Dimensions, and Classifiers—A Review of the State of the Art. *Geocarto Int.* **2021**, *36*, 957–988. [CrossRef]
9. Ball, J.E.; Anderson, D.T.; Chan, C.S. Comprehensive Survey of Deep Learning in Remote Sensing: Theories, Tools, and Challenges for the Community. *J. Appl. Remote Sens.* **2017**, *11*, 1. [CrossRef]
10. Deng, L. A Tutorial Survey of Architectures, Algorithms, and Applications for Deep Learning. *APSIPA Trans. Signal Inf. Process.* **2014**, *3*, e2. [CrossRef]

11. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep Learning in Remote Sensing Applications: A Meta-Analysis and Review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
12. Zhang, L.; Zhang, L.; Du, B. Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40. [CrossRef]
13. Hoeser, T.; Kuenzer, C. Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part I: Evolution and Recent Trends. *Remote Sens.* **2020**, *12*, 1667. [CrossRef]
14. Kattenborn, T.; Leitloff, J.; Schiefer, F.; Hinz, S. Review on Convolutional Neural Networks (CNN) in Vegetation Remote Sensing. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 24–49. [CrossRef]
15. Zang, N.; Cao, Y.; Wang, Y.; Huang, B.; Zhang, L.; Mathiopoulos, P.T. Land-Use Mapping for High-Spatial Resolution Remote Sensing Image Via Deep Learning: A Review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 5372–5391. [CrossRef]
16. Maxwell, A.E.; Warner, T.A.; Guillén, L.A. Accuracy Assessment in Convolutional Neural Network-Based Deep Learning Remote Sensing Studies—Part 2: Recommendations and Best Practices. *Remote Sens.* **2021**, *13*, 2591. [CrossRef]
17. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
18. Flood, N.; Watson, F.; Collett, L. Using a U-Net Convolutional Neural Network to Map Woody Vegetation Extent from High Resolution Satellite Imagery Across Queensland, Australia. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *82*, 101897. [CrossRef]
19. Neupane, B.; Horanont, T.; Hung, N.D. Deep Learning Based Banana Plant Detection and Counting Using High-Resolution Red-Green-Blue (RGB) Images Collected from Unmanned Aerial Vehicle (UAV). *PLoS ONE* **2019**, *14*, e0223906. [CrossRef] [PubMed]
20. Clark, A.; McKechnie, J. Detecting Banana Plantations in the Wet Tropics, Australia, Using Aerial Photography and U-Net. *Appl. Sci.* **2020**, *10*, 2017. [CrossRef]
21. Burke, M.; Driscoll, A.; Lobell, D.B.; Ermon, S. Using Satellite Imagery to Understand and Promote Sustainable Development. *Science* **2021**, *371*, eabe8628. [CrossRef]
22. Zhang, C.; Li, X. Land Use and Land Cover Mapping in the Era of Big Data. *Land* **2022**, *11*, 1692. [CrossRef]
23. Vali, A.; Comai, S.; Matteucci, M. Deep Learning for Land Use and Land Cover Classification Based on Hyperspectral and Multispectral Earth Observation Data: A Review. *Remote Sens.* **2020**, *12*, 2495. [CrossRef]
24. DSITI. *Land Use Summary 1999–2015 for the Atherton Tablelands*; Department of Science, Information Technology and Innovation, Queensland Government: Brisbane, Australia, 2017; p. 26.
25. Dosovitskiy, A.; Springenberg, J.T.; Brox, T. Unsupervised Feature Learning by Augmenting Single Images. *arXiv* **2013**, arXiv:1312.5242.
26. Wieland, M.; Li, Y.; Martinis, S. Multi-Sensor Cloud and Cloud Shadow Segmentation with a Convolutional Neural Network. *Remote Sens. Environ.* **2019**, *230*, 111203. [CrossRef]
27. Sun, Y.; Tian, Y.; Xu, Y. Problems of Encoder-Decoder Frameworks for High-Resolution Remote Sensing Image Segmentation: Structural Stereotype and Insufficient Learning. *Neurocomputing* **2019**, *330*, 297–304. [CrossRef]
28. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv* **2016**, arXiv:1603.04467.
29. Qin, R.; Liu, T. A Review of Landcover Classification with Very-High Resolution Remotely Sensed Optical Images—Analysis Unit, Model Scalability and Transferability. *Remote Sens.* **2022**, *14*, 646. [CrossRef]
30. Van Coillie, F.M.B.; Gardin, S.; Anseel, F.; Duyck, W.; Verbeke, L.P.C.; De Wulf, R.R. Variability of Operator Performance in Remote-Sensing Image Interpretation: The Importance of Human and External Factors. *Int. J. Remote Sens.* **2014**, *35*, 754–778. [CrossRef]
31. Kandel, I.; Castelli, M. The Effect of Batch Size on the Generalizability of the Convolutional Neural Networks on a Histopathology Dataset. *ICT Express* **2020**, *6*, 312–315. [CrossRef]
32. Keskar, N.S.; Mudigere, D.; Nocedal, J.; Smelyanskiy, M.; Tang, P.T.P. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *arXiv* **2017**, arXiv:1609.04836v2.
33. Caye Daudt, R.; Le Saux, B.; Boulch, A.; Gousseau, Y. Multitask Learning for Large-Scale Semantic Change Detection. *Comput. Vis. Image Underst.* **2019**, *187*, 102783. [CrossRef]
34. Liu, S.; Qi, Z.; Li, X.; Yeh, A. Integration of Convolutional Neural Networks and Object-Based Post-Classification Refinement for Land Use and Land Cover Mapping with Optical and SAR Data. *Remote Sens.* **2019**, *11*, 690. [CrossRef]
35. Wurm, M.; Stark, T.; Zhu, X.X.; Weigand, M.; Taubenböck, H. Semantic Segmentation of Slums in Satellite Images Using Transfer Learning on Fully Convolutional Neural Networks. *ISPRS J. Photogramm. Remote Sens.* **2019**, *150*, 59–69. [CrossRef]
36. Stoian, A.; Poulain, V.; Inglada, J.; Poughon, V.; Derksen, D. Land Cover Maps Production with High Resolution Satellite Image Time Series and Convolutional Neural Networks: Adaptations and Limits for Operational Systems. *Remote Sens.* **2019**, *11*, 1986. [CrossRef]