Effective Utilisation of Multiple Open-Source Datasets to Improve Generalisation Performance of Point Cloud Segmentation Models

Abstract—Utilising a single point cloud segmentation model can be desirable in situations where point cloud source, quality, and content is unknown. In these situations the segmentation model must be able to handle these variations with predictable and consistent results. Although deep learning can segment point clouds accurately it often suffers with generalisation, adapting poorly to data which is different than the data it was trained on. To address this issue, we propose to utilise multiple available open source fully annotated datasets to train and test models that are better able to generalise. The open-source datasets we utilise are DublinCity, DALES, ISPRS, Swiss3DCities, SensatUrban, SUM, and H3D [5], [11], [10], [1], [3], [2], [6]. In this paper we discuss the combination of these datasets into a simple training set and challenging test set which evaluates multiple aspects of the generalisation task. We show that a naive combination and training produces improved results as expected. We also show that an improved sampling strategy which decreases sampling variations increases the generalisation performance substantially on top of this. Experiments to find the contributing factor of which variables give this performance boost found that none individually boost performance and rather it is the consistency of samples the model is evaluated on which yields this improvement.

I. INTRODUCTION

Segmentation of a set of unordered 3D points as belonging to a particular object class is a challenging task. Relationships must be found between points to extract and represent local structure and wider context of the scene to accurately classify the class. For example a local flat plane structure could be part of the ground or a building roof depending on the wider elevation context. The task is further complicated when the underlying structure and relationship between points is changed with differences in data collection methods and sensor types. Training models which can handle these changes as well as differences in the scene is important for the practicality of deploying models on new data.

In this paper we do not focus on a particular dataset but rather the combination and testing of a point cloud segmentation model on multiple datasets. We focus this research to point clouds and meshes collected from aerial sensors of outdoor scenes. The datasets contain differences in data collection method, scale, quality, local geography, and building architecture. We combine multiple open source datasets into a cohesive dataset with a training and validation set and a challenging test set.

This paper covers the justification and methodology of how datasets were selected and combined. A naive approach to training a model on the combined data is shown to be effective and produces improved results as expected. Our improved method although simple is effective in increasing the performance further to a more acceptable level. We also perform an ablation of the important factors of model performance on the combined test set to guide focus in future work.

II. RELATED WORK

In this section the datasets utilised in this paper and the research relevant to the methods utilised is covered. The open source labelled aerial datasets are covered first then the methods for semantic segmentation of similar data is covered.

A. Aerial datasets

Table I contains a summary of each identified dataset. The datasets are captured using aerial sensors typically mounted to drones. After data collection this data is compiled into point clouds which span large areas. We note the year of data collection, the sensor type used, the number of classes and the average overhead point density.

The ISPRS dataset [10] is the oldest of this type of dataset being collected in 2012. ISPRS is relatively sparse and small dataset in comparison to the newer datasets. SUM, Dublin city, and SenSatUran [2], [5], [3] datasets contain point clouds from single city urban areas which have been tiled to cover large areas. Swiss3D cities [1] contains point clouds from three cities in Switzerland which have notably different architecture, terrain, and city layouts. Swiss3D cities also includes varying density point clouds of the same area. H3D [6] dataset contains both LIDAR and photogrammetry data of the same area over a number of years resulting in variable quality of dataset.

Classes for these datasets contain ground, vegetation, vehicles, power lines, urban furniture, fences, and buildings. Some datasets further differentiate classes for example splitting vegetation into trees or shrubs and buildings into their component parts such as walls, roof, and windows.

There are datasets of aerial point clouds which we do not consider or run experiments on including LASDU, 3DOMCity, and Bordeaux [14], [7], [12]. 3DOMCity dataset was made utilising a small model of a city block and hence does not cover an extensive area. Utilisation of this dataset could lead to artificially inflated generalisation results due to the smaller nature. Additionally we do not see any conclusions which could be drawn from evaluating on this dataset. Both the LASDU and Bordeaux datasets were not accessible for evaluation and are also excluded from our evaluations.

Although these datasets were collected with similar intentions of developing segmentation models they each have
A summary of publicly available labelled 3D Lidar and Photogrammetry datasets

| Name               | Year   | Type                  | Classes | Point density \(\text{pts/m}^2\) |
|--------------------|--------|-----------------------|---------|---------------------------------|
| DALES [11]         | 2020   | LIDAR                 | 8       | 42.83                           |
| ISPRS-Vaehingen [10]| 2012   | Lidar                 | 9       | 3.45                            |
| LASDU [14]         | 2020   | LIDAR                 | 5       | 1.52                            |
| Bordeaux [12]      | 2020   | LIDAR, Photogrammetry | 5       | 30.21                           |
| Swiss3DCities [1]  | 2021   | Photogrammetry        | 5       | Sparse:1.75 & Medium: 65.343    |
| SenSatUrban [3]    | 2021   | Photogrammetry        | 13      | 261.54                          |
| Dublin City [3]    | 2019   | Lidar                 | 13      | 105.25                          |
| SUM [7]            | 2021   | Photogrammetry        | 6       | 30: 51.07 & 300: 458.13         |
| H3D [6]            | 2016, 2018, 2019 | Lidar,Photogrammetry | 11      | 2016.3: 10.31 & 2018.3: 1025.32 & 2018.11: 1105.32 & 2019.3: 1454.12 |

B. Lidar Segmentation

**Point-based methods** Point-wise multi-layer-perceptron (MLP) methods such as PointNet [8], PointNet++ [9] and RandLaNet [4] use the MLP to create point wise feature representations for the surrounding points. Specifically, PointNet predicts the local point features for each point, the local features are aggregated to a global feature, and finally in the case of semantic segmentation the local and global features are concatenated and passed through a MLP to produce point-wise segmentation. PointNet++ built on this work using hierarchical neural network in order to extract the local features better than PointNet, this involves some sampling and clustering/grouping on the point clouds. The sampling and grouping is computationally expensive and is addressed in the RandLaNet paper. RandLaNet uses random sampling instead of point selection approaches which attempt to find local region centroids. As this process inevitably loses local features they add a module which essentially increases the perceptive field of each point in order to overcome this. RandLaNet performs better on most benchmarks with less computation than PointNet++ and PointNet.

C. Generalisation for Aerial Point Cloud Segmentation

Training and testing point cloud segmentation models on multiple datasets has been researched in the past. The authors of [15] train multiple models on a singular dataset and make predictions on the remaining. In this case the generalisation performance is measures by training on each of DALES and ISPRS and tested on ISPRS and Bordeaux. This train and test split does not provide enough data to identify trends for generalisation.

The authors of [13] utilise two datasets: ISPRS and LASDU. This combination only tests transfer of model performance between the same sensor type with similar datasets. Transferring only to one other dataset may lead to insufficient data to draw generalisation performance conclusions from. In our work we aim to address these short comings by significantly increasing the amount of data that generalisation performance is evaluated on and provide characteristics which can be used to observe performance on sub-problems.

III. Data Preparation

To train a single network across multiple datasets each dataset must be homogenised to a single format. Each point cloud dataset has differences in the format the data is stored to the contents of the dataset such as class labels and data channels from the sensors. In this section we discuss our methodology in the combination of the datasets to enable the generalisation training.

A. Dataset Selection

We aim to select a training and test set that enables us approximate and evaluate the generalisation performance of a particular method. We think that separating a dataset into a simple training set and a more complicated, challenging, and diverse test set allows us to draw such conclusions. Evaluating the complexity of a particular dataset is based on qualitative assessment of the contents and evaluations they may enable. We assess each dataset on content diversity, data collection method, and quality to determine if the dataset forms a challenging benchmark in comparison to others. Additionally we choose a smaller training set in comparison to our test set to provide additional signal to the generalisation performance.

Note: When considering low, medium, and high density point clouds we refer to the ranges of \(\text{pts/m}^2\): \(\text{pts/m}^2 < 2\), \(5 < \text{pts/m}^2 < 100\), \(100 < \text{pts/m}^2\), respectively.

a) Test set: The first choice for test set was the Swiss3DCities dataset. This dataset contains point clouds of three cities Zurich, Davos, and Zug each of which have different urban densities and topography. For example Zug contains a rural town with spaced out buildings, Davos contains point clouds with large elevation changes, and Zurich is the dense urban centre. Additionally we can compare performance on the low and medium density point clouds provided for the same regions.

The H3D dataset contains data over a smaller area than Swiss3DCities but contains data from multiple years as well as photogrammetric mesh and LIDAR scans of the same area. This should allow us to make comparisons between model performance on LIDAR vs photogrammetry datasets by taking out factors such as content. The meshes provided however do not contain a similar detail and density with the 2019 mesh point cloud containing 8.4% of the points the LIDAR counterpart contains. Although the photogrammetry point clouds were generated in conjunction with the LIDAR data these point clouds have much lower quality and detail, see Figure
1. It may be argued that this is in line with photogrammetry quality in comparison to LIDAR but we observe higher quality photogrammetry data in both Sensat and Swiss3DCities. This dataset can also allow for comparisons of the model to perform with data collected at different times at different qualities.

The DublinCity, ISPRS, and Sensat datasets all contain data from a single urban area. The ISPRS dataset is utilised as a test set since it is not large enough to be included as a training set and can provide a measure of generalisation to older point cloud data collection. DublinCity offers a large dataset collected using LIDAR which frequently penetrates through buildings, a phenomenon that doesn’t occur in DALES and H3D nearly as frequently. The Sensat dataset contains data collected using photogrammetry with high detail which in addition covers a diverse cityscape with residential, commercial, industrial, and residential areas. These three datasets offer additional signal to support the generalisation characteristics which may be observed utilising the H3D and Swiss3DCities datasets.

There are some point clouds within the dataset which are not suitable for being combined into the test set and hence are removed or not evaluated. The March 2018 LIDAR and photogrammetry data of H3D dataset does not contain the full area which the March 2016, November 2018, and March 2019 do. Evaluating on this set results in much higher performance as it is a smaller simpler point cloud. March 2016 does not include photogrammetry data and is hence also excluded from evaluation.

b) Training and validation set: The training set simply utilises the remaining datasets: DALES and SUM. These datasets do contain a decent amount of variation which should allow a model to learn the fundamentals to perform well on the test set. The data contains dense urban, rural, and residential areas as well as data collected from both LIDAR and photogrammetry. Using this combination of test and train set we can test generalisation performance on variables such as sensor type, variations in quality, data collection times and methods. The test set also contains data which is not present in the training set like lower quality and density point clouds, different content, and more elevation variation. This should allow for evaluation of the weaknesses and strengths of a segmentation model for generalising across multiple datasets. For the validation set we utilise the same train/val split given in the DALES and SUM papers.

B. Class combinations

Each dataset included in Table I includes different classes, categories, and detail in the labels. In order to combine and train a single model on all datasets these classes need to be combined. In Table II we show that each dataset’s classes can be condensed into three larger categories: ground, buildings, vegetation. Each column shows a dataset and the classes with the label name from their respective paper that have been combined into the rows category. Categories are formed when there are matching classes or sub-classes between all datasets. Street furniture is not labelled accordingly in every dataset and hence does not form a category while buildings are labelled in every dataset. Any classes which remain unmatched into a larger category have labels overwritten as undefined.

In combining the datasets compromises must be made to form the unified set. Despite some datasets having highly detailed labels the resulting combined dataset will contain relatively coarse labels. Although coarse we believe that this combination presents a challenging task when coupled with differences between the dataset contents. In this work we train models utilising the three classes present in Table II to train and test generalisation performance.

C. Dataset preparation

After obtaining all of the datasets we want to put all of them into a format which is consistent to load and process. This involves making all the data the same format, re-issuing the labels for each point and saving them independently. Table II shows the classes which were put into the final dataset.

IV. Model training and evaluation

We do not utilise extra channels such as RGB and intensity in this work to simplify the training methodology and comparisons. Our aim in training and testing the models is not to optimise a particular training strategy but to measure the effects of different training methods on the generalisation
performance. Initially we utilise augmentations and sampling strategies found in the literature and point cloud segmentation implementations. Each model is trained for a maximum of 100 epochs of 500 samples with a controlled batch size of 12. Each training dataset is sampled equally, in our case 250 samples for each dataset per epoch. Classes are weighted by the inverse occurrence in the training data which is approximately $[0.2, 0.4, 0.4]$ for ground, vegetation, and building, respectively. We use a learning rate scheduler which reduces the learning rate by half on the plateau of the validation loss starting at 0.01.

A standard and fixed set of augmentations are applied in all training sequences. All points within a sample are first shuffled for random order so that any structure present is removed. The sample is then shifted such that it is centred and the mean across $x, y, z$ values is zero. Each sample is then augmented with down sampling, rotation around the $z$ axis, and random noise added to points.

We measure model performance using intersection over union (IOU) across the datasets. Where available we break down the performance on subsets of each dataset, e.g. for Swiss3DCities we evaluate performance on each city and density individually. Generalisation values predictable performance over variable data. For this reason our evaluation weights all test sets equally in the performance measure. To evaluate the overall generalisation performance we take the average of the IoU’s over each class, then over the dataset’s subsets, then over each test dataset. We do not present accuracy as IoU is proportional to Accuracy and additionally penalises missing points for a particular class as shown including false positives in the denominator in (2):

$$\text{Accuracy}_c = \frac{TP_c}{TP_c + FN_c + FP_c} \quad (1)$$

$$\text{IoU}_c = \frac{TP_c}{TP_c + FN_c + FP_c} \quad (2)$$

V. POINT CLOUD SAMPLING

Aerial point cloud datasets are typically made up of large blocks of points covering large areas. These blocks contain too much data to process through a model in a single pass. Currently sampling methods of these blocks sample a set number of points the model is capable of processing effectively. The source data is therefore a variable that the model must generalise to. In the case of the datasets utilised in this paper this results in varying the radius and density of each sample.

A. Naive solution

The naive approach takes the standard implementation of sampling from the larger point clouds. This is to take a fixed number, $N$, of points closest to a sampling point. The resulting sample is a spherical encapsulation of surrounding the sample point. We train on each training dataset individually and then on both. The results are shown in Table III.

| Test Set          | DALES | Training set | DALES-SUM |
|-------------------|-------|--------------|-----------|
| DALES             | 0.940 | 0.654        | 0.807     |
| SUM               | 0.5   | 0.728        | 0.623     |
| ISPRS             | 0.811 | 0.566        | 0.713     |
| Dublin            | 0.481 | 0.394        | 0.505     |
| Sensat            | 0.345 | 0.449        | 0.548     |
| H3D all           | 0.471 | 0.444        | 0.527     |
| H3D LIDAR         | 0.503 | 0.460        | 0.599     |
| H3D Mesh          | 0.423 | 0.419        | 0.419     |
| H3D 2016          | 0.778 | 0.415        | 0.656     |
| H3D 2018 LIDAR    | 0.372 | 0.497        | 0.556     |
| H3D 2018 Mesh     | 0.422 | 0.416        | 0.408     |
| H3D 2019 LIDAR    | 0.358 | 0.469        | 0.584     |
| H3D 2019 Mesh     | 0.423 | 0.422        | 0.429     |
| Swiss3DCities All | 0.5204| 0.5246       | 0.5546    |
| Swiss3DCities Sparse | 0.457 | 0.372        | 0.389     |
| Swiss3DCities Medium | 0.553 | 0.585        | 0.612     |
| Swiss3DCities Zurich | 0.509 | 0.57        | 0.553     |
| Swiss3DCities Davos | 0.533 | 0.556        | 0.616     |
| Swiss3DCities Zug | 0.55 | 0.54         | 0.603     |
| Combined LIDAR    | 0.598 | 0.473        | 0.606     |
| Combined Photogrammetry | 0.429 | 0.464       | 0.507     |
| Combined all      | 0.526 | 0.475        | 0.569     |
| Combined all (STD)| 0.173 | 0.069        | 0.083     |

TABLE III. MEAN IOU ACROSS CLASSES GROUND, BUILDING, AND VEGETATION OF SEGMENTATION MODELS. BOLD NUMBERS SIGNIFY HIGHEST PERFORMANCE ON EACH TEST SET.

We include the DALES and SUM performance evaluation in this experiment to show that each model performs well on the individually trained dataset. These values are not included in the combined mean IoU. The performance of the DALES and SUM models on their test sets aligns with the IoU’s seen in their respective papers for the classes included. This indicates that the model is trained to a similar quality to those bench marked in the papers and therefore the generalisation performance can be inferred.

Initially we can see trends emerging where, overall, models trained on LIDAR data alone do not generalise as well to photogrammetry data. This is in contrast to a model trained on photogrammetry data which produced similar performance despite the data collection method. This could be due to the fact that typically photogrammetry data contains much more noise and error than LIDAR. Interestingly the opposite appears
to be true for H3D photogrammetry data for the model trained on DALES.

It is surprising to see that the model performance on Swiss3DCities Davos data of a mountainous region appears to not be as challenging as the dense urban area of Zurich. The H3D photogrammetry data appears to be more difficult for the model trained on the combined set. We can also see that the ISPRS dataset is likely the simplest dataset in the test set consistently performing highest for each model.

The results of training on DALES and SUM datasets individually both clearly show the advantage of the combined training set. In most cases the model trained on the combined dataset has a higher mean IOU than either model trained on the individual datasets. This is a strong signal that the combination of the datasets causes the model to learn better features that are more general to all aerial point clouds. Overall a 8.3% and 19.7% increase in IOU is observed by combining the training sets vs a model only trained on DALES and SUM, respectively. The positive generalisation performance results of this experiment motivate our further exploration.

B. Constant radius

A constant radius sampling attempts to add consistency to the samples which are given to the model. In this case the model will only have to identify larger structures which should remain in the similar overall scale with a maximum density.

It was observed that there was a large disparity between the performance on the sparse and medium subsets of Swiss3D test set. The point density differences between these sets is quite large and with the sampling method would result in between 17m and 109m. This large disparity in sample radius and variable distances between points may confuse the model and produce the observed poor results. To address this we fix the radius of the sample.

Instead of querying the KD-tree up to a number of points we queried up to a set radius. This produces a variable number of points which does not work in batching samples to pass through the model. To address the variable number of points we either randomly drop points or increase the number of points until the desired amount is met. Upsampling in this experiment is done by taking random points in the sample and duplicating them with noise. Upsampled points are then ignored in the evaluation.

Through constraining the radius of the sample it was found that there was between a 0.05 and 0.11 (14.2% and 24.4%) increase in IOU over the DALES-SUM training set. This boost in performance appears to substantially come from increased and more consistent IOU performance on photogrammetry data. The performance on LIDAR data increases at a maximum of 16% while photogrammetry data increases by twice the amount at 32%.

This experiment also inherently tests constant density limits. The maximum density can be calculated as: $d_{\text{max}} = N/(\pi \times r^2)$, where $N$ is the number of points and $r$ is the radius. For example 145m radius fixes the density to be less than 1pt/m$^2$ while 30m fixes the density at 28pt/m$^2$. As most datasets have higher point densities than 1pt/m$^2$ each sample the model is trained and tested on does not contain any upsampling.

C. Constant density

In the previous experiment we observe that there is an advantage for model performance if samples are more consistent. It is unclear if this is the case due to constant radius, more regular densities, or both. Since we must have a fixed number of points a constant density means that a constant radius is also required. In this experiment radius is fixed and density is constant for each model trained. This is accomplished by changing the number of points the model processes. W e chose to train the model using a constant radius of 30m.

Training the model with constant density and radius shows that the RandlaNet model does not continuously improve in performance given higher density data. This can be observed in the datasets which contain higher sampling densities than the ones tested such as Swiss3DCities Medium. We can observe that the down sampled performance does have a marginal effect on the performance but overall plateau at an IOU of 0.746. The performance on each dataset which contains high density data appears to peak or plateau approaching a limit.

We observe a plateau in performance from around 8pts/m$^2$ in which overall performance only degrades on lower density
data. This is a product of the upsampling strategy utilised for our experiments. In the case of Swiss3DCities Sparse subset a density as high as \(2\text{pts/m}^2\) will cause each point to be replicated almost 20 times. This clearly has detrimental effects shown by the decrease in IoU as density increases.

The overall performance of the model on the test sets peaks at \(2\text{pts/m}^2\). It can also be observed that the peak performance on the majority of the test sets is below this density despite containing higher density data. Although we observe fairly small drop off if the density is present this indicates that the model is well equip to handle lower density data. This indicates that the model doesn’t take advantage of the increasing density as would be assumed.

D. Constant radius and density

This experiment attempts to isolate all factors except for the change in radius to observe if there is a particular radius which performs best. To test this we fix the density of the samples to be \(1\text{pts/m}^2\). We achieve this by fixing the density and changing the number of points for each sample. Fixing this value also decreases the number of points the model handles with each sample.

| Test Set       | \(2^{12}/36\) | \(2^{13}/1\) | \(2^{13}/72\) | \(2^{13}/102\) | \(2^{14}/145\) |
|---------------|---------------|---------------|---------------|---------------|---------------|
| ISPRS         | 0.712         | 0.743         | 0.794         | 0.791         | 0.796         |
| Dublin        | 0.685         | 0.686         | 0.740         | 0.710         | 0.708         |
| Sensat        | 0.803         | 0.806         | 0.808         | 0.806         | 0.783         |
| H3D all       | 0.556         | 0.540         | 0.527         | 0.552         | 0.558         |
| H3D LIDAR     | 0.613         | 0.619         | 0.599         | 0.644         | 0.627         |
| H3D Mesh      | 0.613         | 0.614         | 0.604         | 0.614         | 0.613         |
| H3D all       | 0.601         | 0.616         | 0.555         | 0.606         | 0.625         |
| H3D LIDAR     | 0.611         | 0.593         | 0.564         | 0.623         | 0.592         |
| H3D Mesh      | 0.502         | 0.426         | 0.462         | 0.403         | 0.463         |
| H3D 2019 LIDAR| 0.645         | 0.617         | 0.593         | 0.663         | 0.636         |
| H3D 2019 Mesh | 0.532         | 0.444         | 0.475         | 0.425         | 0.463         |
| Swiss3DCities Sparse | 0.671 | 0.652 | 0.624 | 0.589 | 0.526 |
| Swiss3DCities Medium | 0.727 | 0.724 | 0.717 | 0.697 | 0.666 |
| Swiss3DCities Zurich | 0.714 | 0.691 | 0.687 | 0.662 | 0.625 |
| Swiss3DCities Davos | 0.732 | 0.738 | 0.733 | 0.704 | 0.669 |
| Swiss3DCities Zag | 0.671 | 0.674 | 0.656 | 0.663 | 0.666 |
| Combined LIDAR | 0.683 | 0.681 | 0.669 | 0.714 | 0.704 |
| Combined Photogrammetry | 0.676 | 0.639 | 0.653 | 0.628 | 0.624 |
| Combined all   | 0.703 | 0.653 | 0.692 | 0.704 | 0.691 |
| Combined all (STD) | 0.079 | 0.099 | 0.104 | 0.099 | 0.096 |
| PC over naive solution | 123.49% | 121.68% | 121.49% | 123.60% | 121.36% |

We observe that performance of the models do not change substantially across the different radii. Infact we observe that the results are very consistent across the variable radii in these tests. This indicates that it is the consistent density which has the largest impact in performance and not the particular radius. Of course these assumptions may not hold up towards each extreme however we find this to be true within our testing range.

In this experiment the model is trained on between \(2^{12}\) and \(2^{14}\) points. If a smaller radius is more optimal for the model to infer on this performance may be hindered by the lower number of points. We counter this by observing the performance of the models in Table IV with similar radius but higher number of points. Comparing with the constant radius of 30m and 40m the performance is higher with the limited number of points.

### VI. Discussion

In this section we briefly discuss the overall observations we have found through our experimentation. This includes observations of performance on the combined test set as well as the presented sampling strategies.

| Test Set       | Mean IoU | STD | Maximum | Minimum |
|---------------|----------|-----|---------|---------|
| ISPRS         | 0.726    | 0.067 | 0.782 | 0.496 |
| Dublin        | 0.697    | 0.036 | 0.741 | 0.577 |
| Sensat        | 0.811    | 0.019 | 0.840 | 0.781 |
| H3D all       | 0.551    | 0.026 | 0.586 | 0.487 |
| H3D LIDAR     | 0.618    | 0.029 | 0.653 | 0.541 |
| H3D Mesh      | 0.452    | 0.035 | 0.513 | 0.405 |
| H3D 2016      | 0.625    | 0.037 | 0.668 | 0.549 |
| H3D 2018 LIDAR| 0.595    | 0.040 | 0.653 | 0.498 |
| H3D 2018 Mesh | 0.451    | 0.035 | 0.502 | 0.402 |
| H3D 2019 LIDAR| 0.630    | 0.025 | 0.663 | 0.555 |
| H3D 2019 Mesh | 0.456    | 0.036 | 0.523 | 0.406 |

### A. Meta analysis of test set

Through combining the currently available aerial point cloud segmentation datasets we have presented a challenging generalisation baseline. This baseline allows us to draw conclusions on model performance in key characteristics such as sensor type, data quality, and content variability. Performance across these characteristics is important for developing models which can perform reliably and predictably.

Table VII shows an overview of performance of the model on each of the test datasets and their sub-tasks. Overall we observe that the model performs worse on photogrammetry data than LIDAR but this is mainly due to the poor performance on the H3D Mesh data. Disregarding the H3D mesh data produces a higher performance on photogrammetry data however this negates evaluating performance on commonly poorer quality photogrammetry data. We have found that H3D Mesh data is the most challenging dataset while the simplest is the Sensat dataset. We also observe that our sampling strategy has fairly consistent performance across each time period of the H3D dataset. The highest variance in performance belongs to the Swiss3DCities Sparse subset which suffers from the upsampling strategy discussed. Despite handling the density differences between Swiss3DCities Sparse and Medium we still observe discrepancy between performance. Addressing these issues at the model level will be key to generalisation performance.

Throughout the experimentation the largest impact in performance is from model confusion between buildings and vegetation. In the case of photogrammetry vegetation will resemble a shell while LIDAR presents as a mass of points. This discrepancy is most prevalent in a model trained on LIDAR which struggles to perform on photogrammetry data.
B. Sampling strategy observations

We have observed that the sampling strategy used has a large impact on the performance of a model and its ability to generalise across multiple datasets. We first observed that a constant radius sampling strategy improves the results and furthered our experiments to test which attributes affected the performance the most. In our density experimentation from Section V-C we observe that performance plateaus overall but increases if the dataset has density available to it. This leads us to believe that a model which handles these variable densities directly and intently will be able to maintain a much higher generalisation performance.

We have observed that overall a single sample radius did not affect results substantially within our tested ranges. This was against our wider belief that with wider context for each point the model could be more certain of it’s class. It seems that a considerable challenge in this is the models ability to represent local and overall features and hence each sample must contain this information. Although we have not trained utilising either the RGB or intensity channels in this paper that the gains in performance would translate with their inclusion. An extension of this work would focus on utilising any additional channels included in the data.

VII. Conclusion

In this paper we have combined available aerial point cloud datasets to be used to evaluate a segmentation model’s generalisation performance. This includes reasoning about which datasets to use as training/validation/testing, how we combine classes to larger categories, and important characteristics. This combination allows researchers to evaluate performance across different sensor types, contents of datasets, and variability in quality. We have observed that there is a challenge in handling these variables. Rather than handle this at the data level we have shown that a simple sampling strategy can increase performance of the model substantially. We hope that this paper motivates researchers to define this combination as a generalisation benchmark to compare to. Additionally we would also like to see datasets standardise their labels and formatting to aid in additions to the test set in the future.

REFERENCES

[1] G¨ulcan Can, Dario Mantegazza, Gabriele Abbate, S´ebastien Chappuis, and Alessandro Giusti. Semantic segmentation on swiss3dcities: A benchmark study on aerial photogrammetric 3d pointcloud dataset. Pattern Recognition Letters, 150:108–114, 2021. 1, 2
[2] Wexiao Gao, Liangliang Nan, Bas Boom, and Hugo Ledoux. SUM: A benchmark dataset of Semantic Urban Meshes. ISPRS Journal of Photogrammetry and Remote Sensing, 179:108–120, Sept. 2021. 1, 2
[3] Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. Towards semantic segmentation of urban-scale 3d point clouds: A dataset, benchmarks and challenges (cvpr’2021). CVPR 2021, 09 2020. 1, 2
[4] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds. arXiv e-prints, page arXiv:1911.11236, Nov. 2019. 2
[5] S. M. Iman Zolanvari, Susana Ruaano, Aakansha Rana, Alan Cummins, Rogerio Eduardo da Silva, Morteza Rahbar, and Aljosa Smolic. DublinCity: Annotated LiDAR Point Cloud and its Applications. arXiv e-prints, page arXiv:1909.03613, Sept. 2019. 1, 2
[6] Michael K¨olle, Dominik Laupheimer, Stefan Schnohl, Norbert Haala, Franz Rottensteiner, Jan Dirk Wegner, and Hugo Ledoux. The hessisheim 3d (h3d) benchmark on semantic segmentation of high-resolution 3d point clouds and textured meshes from uav lidar and multi-view-stereo. ISPRS Open Journal of Photogrammetry and Remote Sensing, 1:100001, 2021. 1, 2
[7] E. Ozdemir, I. Toschi, and F. Remondino. A Multi-Purpose Benchmark for Photogrammetric Urban 3d Reconstruction in a Controlled Environment. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 4212:53–60, Sept. 2019. 1
[8] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. arXiv e-prints, page arXiv:1612.00953, Dec. 2016. 2
[9] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. arXiv e-prints, page arXiv:1706.02413, June 2017. 2
[10] Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, S´ebastien B´enitez, and U Breitkopf. The isprs benchmark on urban object classification and 3d building reconstruction. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 1-3, 07 2012. 1, 2
[11] Nina Singer and Vijayan Asari. Dales objects: A large scale benchmark database for instance segmentation in aerial lidar. IEEE Access, PP:1–1, 07 2021. 1, 2
[12] I. Toschi, E.M. Farella, M. Welponer, and F. Remondino. Quality-based registration refinement of airborne lidar and photogrammetric point clouds. ISPRS Journal of Photogrammetry and Remote Sensing, 172:160–170, 2021. 1, 2
[13] Xuying Xie, K. Schindler, Jiaojiao Tian, and Xiao Zhu. Exploring cross-city semantic segmentation of als point clouds. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B2-2021:247–254, 06 2021. 2
[14] Zhen Ye, Yusheng Xu, Rong Huang, Xiaohua Tong, Xin Li, Xiangfeng Liu, Kuifeng Luan, Ludwig Hoegner, and Uwe Stilla. Lasdu: A large-scale aerial lidar dataset for semantic labeling in dense urban areas. ISPRS International Journal of Geo-Information, 9(7), 2020. 1, 2
[15] Emre Ozdemir, Fabio Remondino, and Alessandro Golkar. An efficient and general framework for aerial point cloud classification in urban scenarios. Remote Sensing, 13:1985, 05 2021. 2