A GIS Pipeline to Produce GeoAI Datasets from Drone Overhead Imagery

John R. Ballesteros ¹,*, German Sanchez-Torres ² and John W. Branch-Bedoya ¹

¹ Facultad de Minas, Universidad Nacional de Colombia, Medellín 050041, Colombia
² Facultad de Ingeniera, Universidad del Magdalena, Santa Marta 470001, Colombia
* Correspondence: jballes@unal.edu.co

Abstract: Drone imagery is becoming the main source of overhead information to support decisions in many different fields, especially with deep learning integration. Datasets to train object detection and semantic segmentation models to solve geospatial data analysis are called GeoAI datasets. They are composed of images and corresponding labels represented by full-size masks typically obtained by manual digitizing. GIS software is made of a set of tools that can be used to automate tasks using geo-referenced raster and vector layers. This work describes a workflow using GIS tools to produce GeoAI datasets. In particular, it mentions the steps to obtain ground truth data from OSM and use methods for geometric and spectral augmentation and the data fusion of drone imagery. A method semi-automatically produces masks for point and line objects, calculating an optimum buffer distance. Tessellation into chips, pairing and imbalance checking is performed over the image-mask pairs. Dataset splitting into train–validation–test data is done randomly. All of the code for the different methods are provided in the paper, as well as point and road datasets produced as examples of point and line geometries, and the original drone orthomosaic images produced during the research. Semantic segmentation results performed over the point and line datasets using a classical U-Net show that the semi-automatically produced masks, called primitive masks, obtained a higher mIoU compared to other equal-size masks, and almost the same mIoU metric compared to full-size manual masks.

Keywords: GeoAI; GIS; dataset; drone; orthomosaics; U-Net

1. Introduction

Geospatial artificial intelligence, or GeoAI, is an emerging scientific discipline that combines methods in spatial data science and deep learning to extract knowledge from spatial big data. It is an active area of research that has applications in many fields such as disaster management, urban planning, logistics, retail, solar, and many others [1,2]. At the same time, the rapidly increasing availability and quality of drone imagery, the ease of use, and the affordable price of consumer and professional drones are making these technologies converge.

Detection models use rectangle areas that contain objects of interest. Semantic segmentation models make use of full-size masks as labels for objects of interest, or in some cases, such as road center line extraction, uniform (equal-sized) masks are used. There are several tools to support the production of datasets for detection and semantic segmentation in ground imagery. However, the datasets for training GeoAI models are commonly annotated manually, requiring considerable human expert efforts [3]. Furthermore, these datasets for deep learning may suffer from class imbalance or contain an elevated number of misclassified pixels, which means that models may perform poorly, voiding their usability in real applications [4]. This paper presents a GIS pipeline to semi-automatically produce georeferenced datasets for point, line, or polygon objects that can be directly queried from Open Street Maps (OSM) when existing or otherwise digitized over orthomosaics are used...
as base layers. The proposed pipeline describes steps to perform data augmentation and data fusion to robust feature information. A buffer distance parameter is optimized to create appropriate masks after the rasterization of ground truth vector layers, and the resulting datasets are composed of three bands of image–mask chips paired pixel to pixel. The pipeline includes a step to check the data imbalance of image–mask pairs and to produce a Gaussian-like distribution of pixels that guarantees less presence of misclassified pixels.

Robosat, described in [5], is a complete framework to perform GIS layer extraction from satellite orthomosaics and segmentation models such as Yolo V2, U-Net, and PSPNet. However, little is mentioned about drone dataset preparation. Ref. [3] introduces a benchmark over a drone-acquired dataset annotated manually using crowd effort. Another topic of research is to generate synthetic data instead of producing real data. Ref. [6] dealt with data generation for training neural networks in object recognition. Moreover, the researchers in [7] developed an interface to create labels automatically using a generative model for tabular data in medicine.

A specific set of open-source tools for generating and preprocessing geospatial data was created in [8], in which the information is a set of origin/destination points for machine learning algorithm implementation, which differs from the image format addressed in our approach. There are other ways of acquiring data besides drones; ref. [9] proposes a real time geo-spatial data acquisition approach from video streams using deep learning. They considered it a pipeline for the efficient and low-cost approach to geospatial data collection and auto-map generation.

Many authors have tested the performance of different deep learning architectures on drone imagery and confirm the need for a pipeline to produce new ones. For instance, ref. [10] compared four types of deep learning models: GANs, deconvolutional networks, FCNs, and patch-based CNNs. A GAN based on a U-Net model was the second best, with the best F1 score performance on UAV and Google Earth images. Many other researchers have made use of data fusion to enhance model results. Ref. [11] combined UAV ultra-high resolution orthophotos with the digital surface model (DSM) that represents the height of objects to create land cover classification maps. Their experiments demonstrate that the DSM information has greatly enhanced the classification accuracy from 63.93%, with only spectral information, to 94.48% including the DSM. We also use RGB and DSM imagery, but additionally, we incorporate the vegetation atmospheric resistant index (VARI) to test its contribution in the models to segment vehicles and roads.

Imbalance has been tackled directly in model architectures, especially for road datasets and, very infrequently, for point datasets. In our approach, imbalance is studied in the dataset preprocessing point and linear objects. Examples of the first rationale include the Road Vectorization Network (RoadVecNet) presented in [12], which comprises two interlinked U-Net networks to simultaneously perform road segmentation and road vectorization. The authors utilize a loss function called focal loss weighted by median frequency balancing (MFB_FL) to focus on the hard samples, to fix the training data imbalance problem. The researchers in [13] worked out a method of extracting roads and bridges from high-resolution remote sensing images in which edge detection is performed, and the resultant binary edge is vectorized. Their network integrates binary cross entropy to deal with road class imbalance. Ref. [4] presented a weighted balance-loss function over a PSPNet to solve the road class imbalance problem caused by the sparseness of roads. Many scientists concentrate their efforts on modifying the models to improve the resultant masks and not on enriching the input datasets, as in our case. For instance, [14] uses a pre-trained VGG network into a U-Net and attention module to solve road problems such as tortuous shape, connectivity, occlusion, and different scales scenarios.

The rest of the paper is organized as follows. Section 2 deals with the materials and methods used to conduct the research, and develops the theoretical and practical aspects of the proposed pipeline. It contains the description of steps to obtain drone datasets and details on how to include rich, discriminant information. In Section 3, we carry out experimentation and show the results using a buffer distance as a tradeoff
between producing imbalance datasets and datasets enriched with misclassified pixels. The proposed pipeline is tested with the production of two different geometry objects: a road dataset using drone imagery and OSM vector data, and a vehicle dataset obtained from points, similar to the one described in [15]. Section 4 reports the conclusions we drew after experimental analysis. The datasets and the scripts developed for every step of the pipeline are made public and are freely accessible via the GitHub page of this article (https://github.com/DamianoP/DatasetGenerator (accessed on 15 April 2022)). As shown in the Appendix A, all of the code on the repository is open source, licensed under the GNU General Public License v3.0, and freely usable by anyone except for those who require an ArcGIS license.

2. Materials and Methods

Drone imagery is creating new insights in remote sensing thanks to its high spatial resolution, but at the same time, the gathering of new information contained at the centimeter-level demands more robust computer vision algorithms [16].

2.1. GIS Pipeline to Produce GeoAI Datasets

The following pipeline produces datasets to train deep learning models robust to geometric, spectral, and multi-scale variations of geographic objects. The produced datasets consist of image–mask pairs (img, msk), coupled at the pixel level, i.e., drone image chip, binary mask. The datasets are produced separately for point, line, or polygon object geometry. Figure 1 illustrates the steps of the proposed pipeline. Raster layers and vector ground truth data are the input to two separate process lines in which some of the steps are optional depending on the needs of resultant dataset. Next, we describe the different steps of the pipeline.

![Figure 1. GIS pipeline to produce GeoAI datasets from drone imagery.](image)

2.2. Raster Layers: Drone Imagery

Drone imagery is becoming ubiquitous. It is composed of an orthomosaic, a digital surface model (DSM), and a 3D point cloud. Derived products such as the digital terrain model (DTM) can be obtained by post-processing. Orthomosaics are created by stitching images that partially overlap, using a method called Structure from Motion (SfM) [17]. Drone orthomosaics have a very high spatial resolution, measured by the Ground Sample Distance (GSD) [18–20], which is the physical pixel size; a 10 cm GSD means that each pixel in the image has a spatial extent of 10 cm. The GSD of an orthomosaic depends on the altitude of the flight above ground level (AGL) and the camera sensor. Drone photographs are acquired by executing several autonomous flights, using a commercial drone and a controlling application, for example: Dji Phantom 4ProV2 and the Capture App (Professional Photogrammetry and Drone Mapping Software/www.pix4d.com (accessed
on 19 March 2022). Photographs are commonly obtained at heights between 50 and 250 m AGL, depending on the GSD required for the specific application and local flight regulation by the authority (e.g., the FAA regulations). The mapping areas are covered with flight lines using a frontal overlap of 80–85% and a lateral overlap of 70–75%. An orthomosaic to cover one hectare of areal extent is obtained in around one minute of flight at 100 m AGL. Individual images and a GPS log of the flights are processed in a photogrammetric software to obtain default photogrammetric products, which are an orthomosaic, a DSM, and a 3D point cloud of a mapping area. We employed Open Drone Map (www.opendronemap.org (accessed on 9 March 2022)), an open-source software program, to obtain the mentioned products when processing the raw drone images [2]. The WGS1984 is the common geographical coordinate system (GCS) used to geo-reference drone imagery.

Orthomosaics are cropped in two areas: one for the test dataset that is obtained by using parameter $\beta$, which is a percentage (normally 10% to 20%), and the second one for the training and validation datasets using $(1 - \beta)$. Figure 2 illustrates the acquisition and production of drone imagery and how to set aside the orthomosaic area for the test and for the training and validation datasets.

![Figure 2](image-url)

**Figure 2.** (a) Drone imagery workflow, (b) cropping area for training/validation and test datasets: “a” and “b” are the number of pixels horizontally and vertically, respectively, “n” is the number of images in every axis, and “N” is the tessellation size, for instance, $256 \times 256$ pixels. The black area is the residue of tessellation.

### 2.2.1. Geometric Augmentation

Data augmentation improves the performance of deep learning models [21] and model generalization [20, 22–24], at the same time increasing the number of examples to train a model. However, there are not many studies on defining which of the augmentation methods is the best for geographic data. Geometric augmentation consists of transformations in scale, angle, and form of images. These variations depend on the field of application and, particularly, on the requirements imposed on a model. For instance, ninety-degree mirroring may not be applicable to common objects such as dogs or bikes, but they are to overhead imagery. The most important geometric augmentation methods for geographic objects are [21]:

- **Rotation:** consists of small clockwise rotations of images; suggested value is 10 degrees [22].
- **Mirroring:** a transformation in which the upper and lower, or right and left, parts of images interchange position. They are commonly referred to as vertical and horizontal mirroring.
- **Resizing or zooming:** the magnification of certain parts of an image, zooming in or out.
- **Cropping:** the trimming of an image at certain place.
• Deformation: the elastic change of the proportion of image dimensions. It is a common phenomenon that occurs in the borders of orthomosaics [17].
• Overlapping: the repetition of a part of an image measured by a percentage (%).

2.2.2. Spectral Augmentation

Spectral augmentation is the change in brightness, contrast, and intensity (gamma value) of images [21]. Typically, a 10% increment or decrement of the current values is applied. They are described as follows:

• Brightness: the amount of light in an image. It increases the overall lightness of the image—for example, making dark colors lighter, and light colors whiter (GIS Mapping Software, Location Intelligence and Spatial Analytics | Esri, www.esri.com (accessed on 2 May 2022))
• Contrast: the difference between the darkest and lightest colors of an image. An adjustment of contrast may result in a crisper image, making image features become easier to distinguish (GIS Mapping Software, Location Intelligence and Spatial Analytics | Esri, www.esri.com (accessed on 2 May 2022)).
• Intensity or gamma value: refers to the degree of contrast between the mid-level gray values of an image. It does not change the extreme pixel values, the black or white—it only affects the middle values [21]. A gamma correction controls the brightness of an image. Gamma values lower than one decrease the contrast in the darker areas and increase it in the lighter areas. It changes the image without saturating the dark or light areas, and doing this brings out the details of lighter features, such as building roofs. On the other hand, gamma values greater than one increase the contrast in darker areas, such as shadows from buildings or trees in roads. They also help bring out details in lower elevation areas when working with elevation data such as DSM or DTM. Gamma can modify the brightness, but also the ratios of red to green to blue (GIS Mapping Software, Location Intelligence and Spatial Analytics | Esri, www.esri.com (accessed on 2 May 2022)).

2.2.3. Data Fusion

Due to computational limitations, most deep learning models for computer vision make use of images with three channels, i.e., RGB images [25]. Data fusion is a way to incorporate additional discriminant information to the available channels. Object height can be a discriminant variable where intricate spatial relations may exist. For instance, the spatial relations between vehicles, roads, trees, and buildings are good examples of such a case. There are also many popular vegetation indexes developed in remote sensing and are mostly used in agricultural monitoring. The well-known Normalized Difference Vegetation Index (NDVI) quantifies the health of vegetation by measuring the difference between bands in a near-infrared image (NIR) [26]. Data fusion can be used for integrating height or indexes into a dataset, as follows:

• Height: the DSM, which contains the height of objects in an image, can be fused with the orthomosaics by adding it either algebraically or logarithmically to each red (R), green (G), and blue (B) band as stated in (1) and (2). Another option is replacing any of the bands with DSM, as in (3).

\[
HRGB = (R + DSM), (G + DSM), (B + DSM) \quad (1)
\]
\[
HLRGB = (R + Log(DSM)), (G + Log(DSM)), (G + Log(DSM)) \quad (2)
\]
\[
RGDSM = (R), (G), (DSM) \quad (3)
\]
In any case, the resultant image is a three-band false color composite [27] with values ranging between 0 and 255, so values of every band should be rescaled to that interval [11] using Equation (4).

\[
\text{Rescaled band} = \left( \text{pxval} - \text{minpxval} \right) \times 255 / \left( \text{maxpxval} - \text{minpxval} \right)
\]

where \( \text{minpxval} \) and \( \text{maxpxval} \) are the minimum and maximum values of the band, respectively. More datasets now include height as a way to improve image understanding, for example, the NYU depth V2, the SUN RGB-D, and the HAGDAVS [15,28].

- Index: may replace one of the RGB bands of a drone orthomosaic with the values of an index. The Visible Atmospherically Resistant Index (VARI) was developed by [29], based on a measurement of corn and soybean crops in the midwestern United States, to estimate the fraction of vegetation in a scene, with low sensitivity to atmospheric effects in the visible portion of the spectrum. It is exactly what occurs in low-altitude drone imagery [26]. Equation (5) allows the calculation of the VARI for an orthomosaic using the red, green, and blue bands of an image.

\[
\text{VARI} = \frac{\text{Green} - \text{Red}}{\text{Green} + \text{Red} - \text{Blue}}
\]

The VARI should also be rescaled to the orthomosaics’ interval values [0, 255] using (4), obtaining the NVARI.

2.3. Vector Layers: Ground Truth

Ground truth data are obtained by querying OSM vector layers using a Python script and “overpass” open-source library (https://pypi.org/project/overpass/ (accessed on 2 March 2022)). Depending on the part of the world, roads, pois, rivers, and, less frequently, buildings can be downloaded in a matter of seconds. Appendix A links a repository for the Python scripts and data used in this paper. Specific data of interest that cannot be found in OSM, for example, vehicles, people, and animals, should be digitalized in the screen from the start. This is done by manual tracing point, line, and polygon to represent objects using drone orthomosaics as base geo-referenced layers. Point objects are those than can be depicted as \((x, y)\) coordinates at a geographical extent. Line objects are those in which length is a lot larger than width. They are digitalized by adding vertices \((x, y)\) at any change of direction and have at least two vertices. Polygon objects are regions, and vertices are created at every change of direction until the last vertex coincides with the initial one.

Vector Masks, Raster Masks, and Color Masks

The ground truth point, line, or polygon layers are buffered using a distance parameter, obtaining a vector mask of objects of interest, without the need for manual digitalization. The buffer distance is typically measured from the center point or line and used to increase the size of the point and line vector geometries, with the aim of reducing the imbalance of their vector masks. The buffer distance is a “tradeoff” between obtaining imbalanced thin masks with no misclassifications, and wider masks with more pixels of mixed classes. Polygon object masks are less affected by imbalance, thus, the buffer distance used is zero (0). So, the problem is reduced to find the optimum distance to produce the point and line masks. Once this value is calculated, vector masks are converted to raster (raster masks) to produce an image, with the same extension and coordinate system of the base orthomosaic. We call a raster mask produced in such a way a “primitive mask”. Primitive masks can be binary (black and white), and represent only one object of interest (positive class) and its background (negative class). The positive class is encoded in white (class = 1) and competes against the ground, dominant class, encoded as black (class = 0). A color raster mask is used when extracting object attributes, for instance, road speed, vehicle type, roof material, and many others. Figure 3 shows an example of a manually produced full-size mask, and an equal size mask obtained by buffering roads in drone imagery. Full-size masks are
generally less imbalanced than equal-sized ones, but extracting the road centerline from them is more complex.

Figure 3. Type of masks. (a) Full-size and equal-size binary masks, and (b) color mask by road type.

2.4. Image Tessellation, Imbalance Check, Pairing, and Splitting

Due to computational restrictions, it is common to train deep learning models with square 256 × 256 px images. In this respect, orthomosaics and raster masks (binary or color) are huge; thus, they should be tessellated at a desired size \( N \), producing \( (N \times N) \) image chips, for example, 256 × 256 pixels. Since many geographical objects are scarce with respect to the ground, they produce an imbalanced mask. Class imbalance is a common problem that affects the performance of deep learning models, moving the decision boundary towards the dominant class [30]. The imbalance of the positive class can be calculated for a specific dataset with \( n \) images, as in (6).

\[
\text{Imbalance of positive class} = \sum_{i=1}^{n} \frac{\text{pixels of positive class}}{\text{pixels of positive class} + \text{pixels of negative class}} \tag{6}
\]

Values of around 0.5 in (6) correspond to a perfect pixel balanced mask, and values below 0.01 are an extremely imbalanced mask. Instead of calculating the imbalance on the whole raster mask, an imbalance check may be applied on every mask using a threshold \( t \). A proper value of the \( t \) parameter should be chosen depending on the specific dataset and the geometry of the objects. A very small value of \( t (<0.01) \) is equivalent to keeping the original dataset unchanged. In a similar way, a high value of \( t (>0.1) \) may restrict the model to be tested in hard cases. After that, every image–mask pair corresponding to a balance mask is saved as a whole image of \( (2N \times N) \) pixels, for instance 512 × 256 pixels. Finally, random splitting into training and validation datasets is done using a proportion: \((1 \rightarrow \alpha)\) for training, and \(\alpha\) for validation.

3. Results

To test the pipeline, two datasets were produced. The first one is a vehicle dataset, which is represented by point geometry. Vehicles are traced manually as points in ArcGIS software over the drone imagery. The second is a road dataset represented by line geometry. Vector roads GT were queried from OSM and converted to shapefile format using a Python script (Appendix A). The drone imagery used was acquired for five small settlements in Colombia, South America. Figure 4 shows an example of the acquired drone imagery. Table 1 presents the metadata of the drone imagery, where Lonmin, Lonmax and Latmin, Latmax are the minimum and maximum longitude and latitude in decimal degrees of the orthomosaics extent, respectively.
Table 1. Drone imagery metadata.

| Settlement          | Geographic Extent (Lonmin; Latmin) (Lonmax; Latmax) | Flight Height (m) | GSD (cm/px) | Number of Pixels | Area (Hectares) |
|---------------------|-----------------------------------------------------|-------------------|-------------|------------------|-----------------|
| El Retiro, Ant.     | (−75.5075858485094; 6.05456672000301)              | 120               | 7           | 11,276, 16,914   | 82.9            |
|                     | (−75.4995986448169; 6.06544416605448)             |                   |             |                  |                 |
|                     | (−75.4579001836735; 6.03130980894862)             |                   |             |                  |                 |
| La Ceja, Ant.       | (−75.4332962779884; 6.0342695019348)              |                   |             |                  |                 |
| Prado_largo, Ant.   | (−75.531188383421; 6.15636546472326)              | 90                | 5.7         | 20,826, 16,829   | 40              |
|                     | (−75.5226877620765; 6.16018600622437)             |                   |             |                  |                 |
| Rionegro, Ant.      | (−75.3809074659528; 6.13947401033623)             | 80                | 5.5         | 8847, 18,895     | 62.7            |
|                     | (−75.3760197352806; 6.14988050247727)             |                   |             |                  |                 |
|                     | (−75.8715703129762; 6.1787682641961)             | 180               | 8.6         | 20,744, 30,428   | 572.0           |

3.1. Method for Producing Primitive Masks

A mask for a specific object should contain as many pixels as possible that belong to the object of interest, and at the same time, the least number of misclassified pixels possible. Attending to that, one way to calculate the optimum buffer distance of a mask is to graph the standard deviation of the pixel values of every band of orthomosaics vs. the mask buffer distance for an intended dataset. We created a vehicle and a road dataset with differently sized masks, starting with 50 cm and increasing by 50 cm until 3 m wide masks were created, and the standard deviation of the pixel values was calculated. Figure 5 illustrates the resulting graph of the standard deviation of the pixel vs. buffer distance for the road dataset.

Figure 5. Primitive mask and the optimum value of the buffer distance for the road dataset.
In the graph of Figure 5, for a buffer distance of 100 cm (orange vertical line), there is practically no change in the standard deviation of the RGB value distribution, which seems to indicate that 1 m is the distance with the best Gaussian-like distribution of RGB values, and so this is the buffer distance of the primitive mask for this dataset. The blue band distribution suggests that replacing the blue channel with DSM does not seem to work as well as adding DSM to every band. Figure 6 shows how the buffer distance affects the distribution of RGB pixel values of roads.

Figure 6. Cont.
Figure 6. RGB pixel value distributions vs. mask buffer distance for roads dataset. (a) RGB distribution for the whole orthomosaic, imbalance = 8.03%, (b) 50 cm buffer distance, imbalance = 2.13%, (c) 1 m buffer distance, imbalance = 4.35%, (d) 2 m, imbalance = 7.86%, (e) 3 m, imbalance = 13.56%, (f) full-size mask, imbalance = 11.63%. The same experiment was performed over all orthomosaics, and the values are probably similar because the roads are of similar size and materials in the mapping zone.

We also created different size masks for vehicles using distances of 50 cm to 150 cm and compared the pixel distributions vs. the buffer distance and the pixel distribution of the full-size masks. Figure 7 shows the pixel distribution of all masks for all orthomosaics, and the graph to obtain the primitive mask for the vehicle dataset.
Figure 7. Cont.
Figure 7. Primitive mask for vehicle dataset. (a) 50 cm pixel distribution and mask, (b) 100 cm pixel distribution and mask, (c) 150 cm pixel distribution and mask, (d) full-size pixel distribution and mask, (e) graph of standard deviation of RGB pixel values vs. buffer distance.

As can be observed in Figure 7, compared to the road masks, the vehicle masks do not exhibit a perfectly Gaussian-shaped curve, probably because vehicles are not uniformly colored. Although full-size masks consist of more pixels, the RBG distribution looks very similar to the distribution of other distance masks; moreover, full-size masks have a slightly higher standard deviation (brown vertical line) than the 1 m buffer mask (orange vertical line). The graph of the standard deviation vs. distance shows that a buffer distance of 100 cm seems to be the most appropriate buffer distance for producing primitive masks in roads.

3.2. Dataset Production

All of the proposed geometric and spectral augmentation methods are applied to both example datasets. An overlap of 20% is suggested. Increasing angle rotations of 10 degrees clockwise are used, as well as mirroring (90 and 180 degrees). Appendix A contains the link to our implementation of data augmentation in Jupyter Notebooks. Figure 8 shows examples of (img, msk) pairs obtained by data fusion. Figure 8 presents an example of RGDSM and RVARIB false color composite images obtained by data fusion.

Figure 8. Data fusion imagery. RGDSM (left), RVARIB (right).

Different sizes for tessellations can be used, for instance, $256 \times 256$, $512 \times 512$, and $1024 \times 1024$ pixels. The images and corresponding masks are then paired into single images (img, msk) with sizes of $512 \times 256$, $1024 \times 512$, and $2048 \times 1024$ pixels, respectively. Every (img, msk) pair is checked to pass an imbalance threshold chosen by the user, for example, 1%, 5%, or 10%. Vehicles and roads pixels are imbalanced with respect to the background. Figure 9 shows an example of vehicle and road datasets produced with the pipeline. Appendix A contains the link to download these datasets.
3.3. Dataset Evaluation

We trained a standard U-Net segmentation model [31] with masks of different buffer distances for example vehicle and road datasets, and compared the results using the mIoU metric to account for model learning of the geometric aspect of geographic objects [17]. Figure 10 exhibits the mIoU results obtained with the U-Net. For both datasets, a buffer distance of 1 m produces the second-best mIoU results after the biggest buffer distance used. However, the road structure and vehicle position are easier to extract from a thinner mask, and furthermore, thinner masks have a smaller number of misclassified pixels of other classes such as buildings and trees, which can also cause a problem if using multiclass masks for segmentation.
Figure 10. Segmentation results mIoU vs. mask buffer size. (a) Vehicles, (b) roads.
For the vehicle dataset, the graph of mIoU vs. buffer size and the qualitative segmentation results show that the semi-automatic datasets with buffer distances of 100, 150, 200, and 300 cm surpass the mIoU value of the full-size masks (mIoU = 0.455). However, primitive masks (100 cm) have an inferior mIoU value compared to the masks of 250 cm.

For the road dataset, the graph of mIoU vs. buffer size and the segmentation results show that primitive masks (100 cm) slightly exceed the mIoU value of the full-size masks (mIoU = 0.595). Again, the primitive masks have an inferior mIoU value compared to the masks of 500 cm. In both cases, the extremely imbalanced datasets (threshold < 1%) obtained with a buffer distance of 50 cm did not generate or barely generated any segmentation results. All of the datasets, independent of the buffer distance used, exhibited discontinuities (false-negative pixels) and irregularities (false-positive pixels) in the resultant masks.

The use of 90-degree mirroring data augmentation and data fusion for the road dataset increased the model performance. Figure 11 and Table 2 exhibit these results using a buffer distance of 100 cm. It seems that including the height of objects is more effective than both using the VARI index and combining VARI index and height in the road dataset.

![Figure 11. (a) RGDSM, (b) RVARIB, (c) HRGB, (d) HRVARIB.](image)

Table 2. Road dataset performance using data fusion.

| Dataset | mIoU |
|---------|------|
| RGDSM   | 0.725|
| RVARIB  | 0.549|
| HRGB    | 0.621|
| HRVARIB | 0.508|

4. Conclusions

This pipeline allows the creation of datasets in a semi-automatic way and enables the inclusion of highly discriminant characteristics of objects of interest by performing height data fusion, index, geometric, and spectral augmentation.

Dataset imbalance is closely related to model performance; for instance, using a buffer distance of 50 cm produced imbalance values of around 1% for vehicles and 2% for roads. These masks did not generate segmentation results for either vehicle or road datasets using the U-Net.

The results show that primitive masks can be used as a replacement for full-size masks for the used point and line example datasets, without sacrificing performance. Choosing a larger buffer distance improved the metric, but contaminated the training masks by adding pixels of other object classes, but at the same time, it reduced the imbalance of the vehicle and road datasets. The higher values of mIoU obtained for a larger buffer distance seem to show that misclassified pixels are less important than imbalance classes for the U-Net.
model. Different combinations of data augmentation and false composite images can be performed within the pipeline, with the results showing that including the height of objects improves model performance. However, more research is needed in the use of VARI to help in discriminating objects. The proposed pipeline supports the semi-automatic production of different datasets to investigate those relationships. The production of multi-color mask datasets was not tested in the study. The use of the pipeline with satellite imagery is proposed as a future research approach.

A limitation of the proposed pipeline is that the user needs to choose different parameters, such as the buffer distance, the imbalance threshold, and the splitting percentage of the datasets to be produced; however, default values are suggested.

**Author Contributions:** Conceptualization, John R. Ballesteros, German Sanchez-Torres, and John W. Branch-Bedoya; methodology, John R. Ballesteros, German Sanchez-Torres, and John W. Branch-Bedoya; software, John R. Ballesteros; validation, John R. Ballesteros, German Sanchez-Torres, and John W. Branch-Bedoya; investigation, John R. Ballesteros; resources, John W. Branch-Bedoya; data curation, John R. Ballesteros and German Sanchez-Torres; writing—original draft preparation, John R. Ballesteros; writing—review and editing, German Sanchez-Torres; visualization, John R. Ballesteros and German Sanchez-Torres; supervision, John W. Branch-Bedoya. 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:** Not applicable.

**Acknowledgments:** We would like to thank GIDIA (Research Group in Artificial Intelligence of the Universidad Nacional de Colombia) for providing guidelines, reviews of the work, and the use of facilities.

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

**Appendix A**

Code: https://github.com/jrballesteros/GeoAI_Datasets (accessed on 20 September 2022).

Domino dataset: https://zenodo.org/record/5718809 (accessed on 20 September 2022).

Drone Road dataset: https://zenodo.org/record/7020196 (accessed on 20 September 2022).

Drone Orthomosaics and Raster Masks: https://zenodo.org/record/7019074 (accessed on 20 September 2022).

HAGDAVS dataset: https://zenodo.org/record/6323712 (accessed on 20 September 2022).

**References**

1. Song, Y.; Huang, B.; Cai, J.; Chen, B. Dynamic Assessments of Population Exposure to Urban Greenspace Using Multi-Source Big Data. *Sci. Total Environ.* **2018**, *634*, 1315–1325. [CrossRef]

2. Ballesteros, J.R.; Sanchez-Torres, G.; Branch, J.W. Automatic Road Extraction in Small Urban Areas of Developing Countries Using Drone Imagery and Image Translation. In Proceedings of the 2021 2nd Sustainable Cities Latin America Conference (SCLA), Online, 25–27 August 2021; pp. 1–6.

3. Vanschoren, J. Aerial Imagery Pixel-Level Segmentation Aerial Imagery Pixel-Level Segmentation. Available online: https://www.semanticscholar.org/paper/Aerial-Imagery-Pixel-level-Segmentation-Aerial-Vanschoren/7dadc3aff6f2b49282c06a2aa6effbd4267 (accessed on 26 February 2022).

4. Gao, X.; Sun, X.; Zhang, Y.; Yan, M.; Xu, G.; Sun, H.; Jiao, J.; Fu, K. An End-to-End Neural Network for Road Extraction From Remote Sensing Imagery by Multiple Feature Pyramid Network. *IEEE Access* **2018**, *6*, 39401–39414. [CrossRef]

5. Ng, V.; Hofmann, D. Scalable Feature Extraction with Aerial and Satellite Imagery. In Proceedings of the 17th Python in Science Conference (SCIPY 2018), Austin, TX, USA, 9–15 July 2018; pp. 145–151.

6. Perri, D.; Simonetti, M.; Gervasi, O. Synthetic Data Generation to Speed-Up the Object Recognition Pipeline. *Electronics* **2022**, *11*, 2. [CrossRef]
7. Ratner, A.; Bach, S.H.; Ehrenberg, H.; Fries, J.; Wu, S.; Ré, C. Snorkel: Rapid Training Data Creation with Weak Supervision. In Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases, Munich, Germany, 28 August–1 September 2017; Volume 11, p. 269. [CrossRef]

8. Golubev, A.; Chechetkin, I.; Parygin, D.; Sokolov, A.; Scherbakov, M. Geospatial Data Generation and Preprocessing Tools for Urban Computing System Development. *Procedia Comput. Sci.*, 2016, 101, 217–226. [CrossRef]

9. Al-Azizi, J.I.; Shafri, H.Z.M.; Hashim, S.J.B.; Mansor, S.B. DeepAutoMapping: Low-Cost and Real-Time Geospatial Map Generation Method Using Deep Learning and Video Streams. *Earth Sci. Inf.*, 2020, 15, 1481–1494. [CrossRef]

10. Abdollahi, A.; Pradhan, B.; Shukla, N.; Chakraborty, S.; Alamri, A. Deep Learning Approaches Applied to Remote Sensing Datasets for Road Extraction: A State-Of-The-Art Review. *Remote Sens.*, 2020, 12, 1441. [CrossRef]

11. Zhang, Q.; Qin, R.; Huang, X.; Fang, Y.; Liu, L. Classification of Ultra-High Resolution Orthophotos Combined with DSM Using a Dual Morphological Top Hat Profile. *Remote Sens.* 2015, 7, 16422–16440. [CrossRef]

12. Abdollahi, A.; Pradhan, B.; Alamri, A. RoadVecNet: A New Approach for Simultaneous Road Network Segmentation and Vectorization from Aerial and Google Earth Imagery in a Complex Urban Set-Up. *GISci. Remote Sens.* 2021, 58, 1151–1174. [CrossRef]

13. Yang, W.; Gao, X.; Zhang, C.; Tong, F.; Chen, G.; Xiao, Z. Bridge Extraction Algorithm Based on Deep Learning and High-Resolution Satellite Image. *Sci. Program.*, 2021, e9961963. [CrossRef]

14. Gong, Z.; Xu, L.; Tian, Z.; Bao, J.; Ming, D. Road Network Extraction and Vectorization of Remote Sensing Images Based on Deep Learning. In Proceedings of the 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (IToEC), Chongqing, China, 12–14 June 2020; pp. 303–307.

15. Ballesteros, J.R.; Sanchez-Torres, G.; Branch-Bedoya, J.W. HAGDAVS: Height-Augmented Geo-Located Dataset for Detection and Semantic Segmentation of Vehicles in Drone Aerial Orthomosaics. *Data* 2022, 7, 50. [CrossRef]

16. Avola, D.; Pannone, D. MAGi: Multistream Aerial Segmentation of Ground Images with Small-Scale Drones. *Drones* 2021, 5, 111. [CrossRef]

17. Kameyama, S.; Sugiiura, K. Effects of Differences in Structure from Motion Software on Image Processing of Unmanned Aerial Vehicle Photography and Estimation of Crown Area and Tree Height in Forests. *Remote Sens.* 2021, 13, 626. [CrossRef]

18. Heffels, M.; Vanschoren, J. Aerial Imagery Pixel-Level Segmentation. *arXiv 2020*, arXiv:2012.02024.

19. Shermeyer, J.; Etten, A. The Effects of Super-Resolution on Object Detection Performance in Satellite Imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 1432–1441.

20. Weir, N.; Lindenbaum, D.; Bastidas, A.; Etten, A.; Kumar, V.; Mcpherson, S.; Shermeyer, J.; Tang, H. SpaceNet MVOI: A Multi-View Overhead Imagery Dataset. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 992–1001.

21. Buslaev, A.; Iglovikov, VI.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and Flexible Image Augmentations. *Information* 2020, 11, 125. [CrossRef]

22. Blaga, B.-C.-Z.; Nedevschi, S. A Critical Evaluation of Aerial Datasets for Semantic Segmentation. In Proceedings of the 2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania, 3–5 September 2020; pp. 353–360.

23. Long, Y.; Xia, G.-S.; Li, S.; Yang, W.; Yang, M.Y.; Zhu, X.X.; Zhang, L.; Li, D. On Creating Benchmark Dataset for Aerial Image Interpretation: Reviews, Guidelines, and Million-AID. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 2021, 14, 4205–4230. [CrossRef]

24. Song, A.; Kim, Y. Semantic Segmentation of Remote-Sensing Imagery Using Heterogeneous Big Data: International Society for Photogrammetry and Remote Sensing Potsdam and Cityscape Datasets. *ISPRS Int. J. Geo-Inf.* 2020, 9, 601. [CrossRef]

25. Xu, Y.; Xie, Z.; Feng, Y.; Chen, Z. Road Extraction from High-Resolution Remote Sensing Imagery Using Deep Learning. *Remote Sens.* 2018, 10, 1461. [CrossRef]

26. Eng, L.S.; Ismail, R.; Hashim, W.; Baharum, A. The Use of VARI, GLI, and Vilgreen Formulas in Detecting Vegetation In Aerial Images. *IITech 2019*, 10, 1385. [CrossRef]

27. López-Tapia, S.; Ruiz, P.; Smith, M.; Matthews, J.; Zercher, B.; Sydoreenko, L.; Varia, N.; Jin, Y.; Wang, M.; Dunn, J.B.; et al. Machine Learning with High-Resolution Aerial Imagery and Data Fusion to Improve and Automate the Detection of Wetlands. *Int. J. Appl. Earth Obs. Geoinf.* 2021, 105, 102581. [CrossRef]

28. Sun, W.; Wang, R. Fully Convolutional Networks for Semantic Segmentation of Very High Resolution Remotely Sensed Images Combined With DSM. *IEEE Geosci. Remote Sens. Lett.* 2018, 15, 474–478. [CrossRef]

29. Gitelson, A.A.; Stark, R.; Grits, U.; Rundquist, D.; Kaufman, Y.; Derry, D. Vegetation and Soil Lines in Visible Spectral Space: A Concept and Technique for Remote Estimation of Vegetation Fraction. *Int. J. Remote Sens.* 2002, 23, 2537–2562. [CrossRef]

30. Wang, S.; Liu, W.; Wu, J.; Cao, L.; Meng, Q.; Kennedy, P.J. Training Deep Neural Networks on Imbalanced Data Sets. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 4368–4374.

31. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.