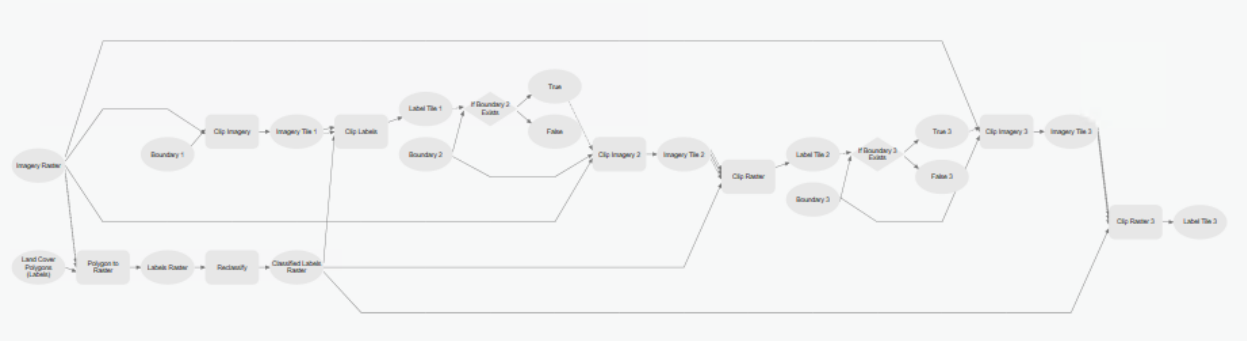
Tutorial 1: Data Preparation for Deep Learning

A core component of the deep learning process is training the model with labelled data. For our purposes, training data is a raster file in .tif format that is exported from ArcGIS Pro. When training our deep learning model, we have a choice to either **create new training data** or **obtain existing training data**. This tutorial will explain the process of preparing existing data for training, focusing specifically on using past SAL projects or other GIS resources to train the model.

Since this tutorial involves a lot of geoprocessing tasks in ArcGIS, I have created a model that automates almost every step in this tutorial. To use it, simply download the model from the GitLab repository for this project and open it in an ArcGIS Pro. You will still need to create 3 clip extent polygons, and specify the reclassification instructions as detailed in this tutorial.

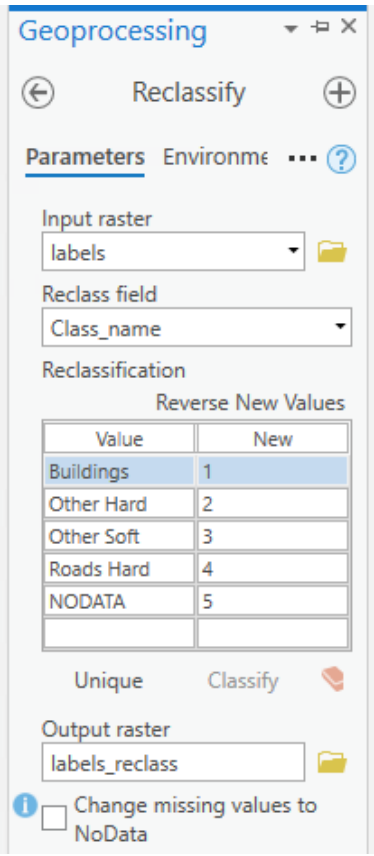


You may also choose not to use the model if you are working with a small amount of data, but preparing large amounts of data can be tedious if done manually.

When training the model, it is essential that we have two sets of data, **imagery** and **labels**. The framework will train the model by matching the imagery and labels together.

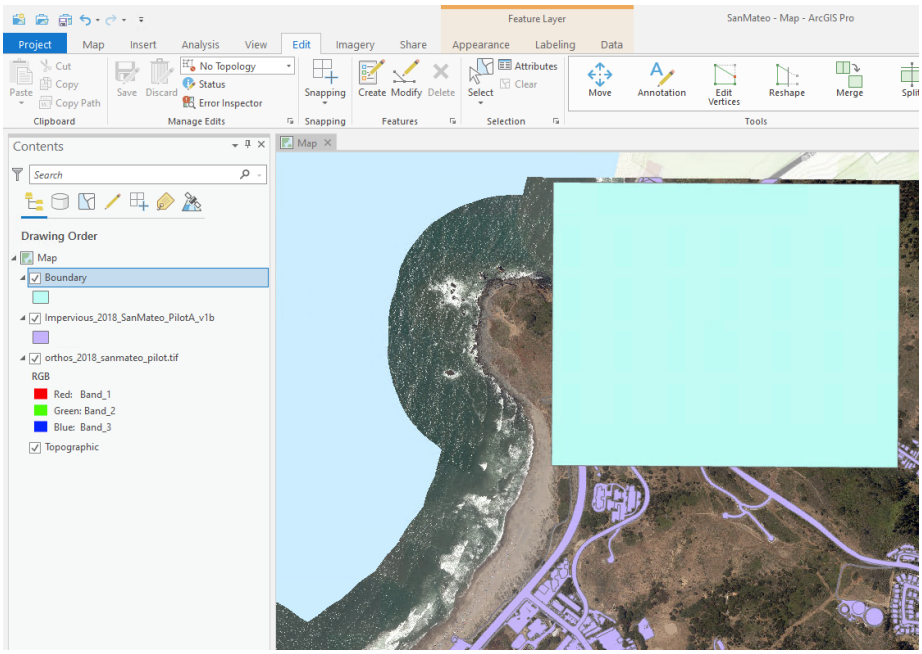
Oftentimes, the labels will be in vector format. To convert the labels to raster format, use the Polygon to Raster tool.

1. If your labels are in vector format, use the Polygon to Raster tool to convert them to raster format. Use your raster imagery as input for the cell size and snap raster fields to ensure that the imagery and labels align perfectly.
2. Once the labels are rasterized, run the reclassify tool to classify the fields you will be using in your model. For instance, if you have land cover data that classifies tree canopy, water, and impervious surface, but you are only interested in impervious surfaces, you can reclassify your raster to only contain the necessary information. You should also reclassify all of the NODATA values as a unique class, as shown below.



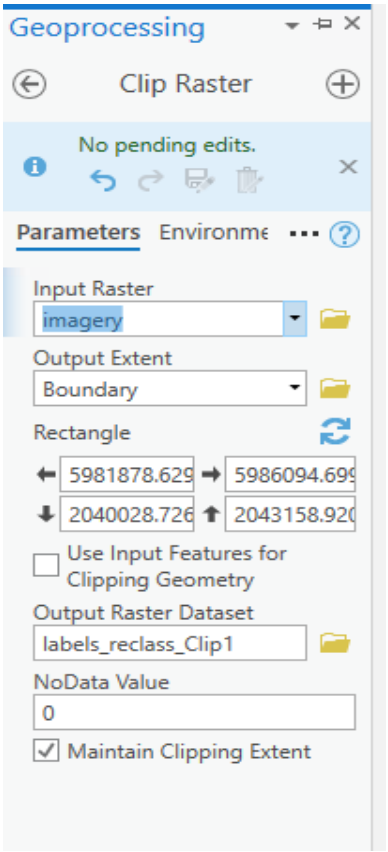
Now that you have imagery and labels in raster format, it is time to clip the data into manageable tiles for training.

3. Use the Create Feature Class tool to create a polygon feature class.
4. Draw a polygon in this feature class using Create Features in the edit menu. Use the rectangle tool to create a perfectly rectangular tile around the area you want to use for training data. I tend to keep each tile about 4000 x 4000 pixels to avoid large file sizes.



Next, we will clip the imagery to the boundary box created in the last step.

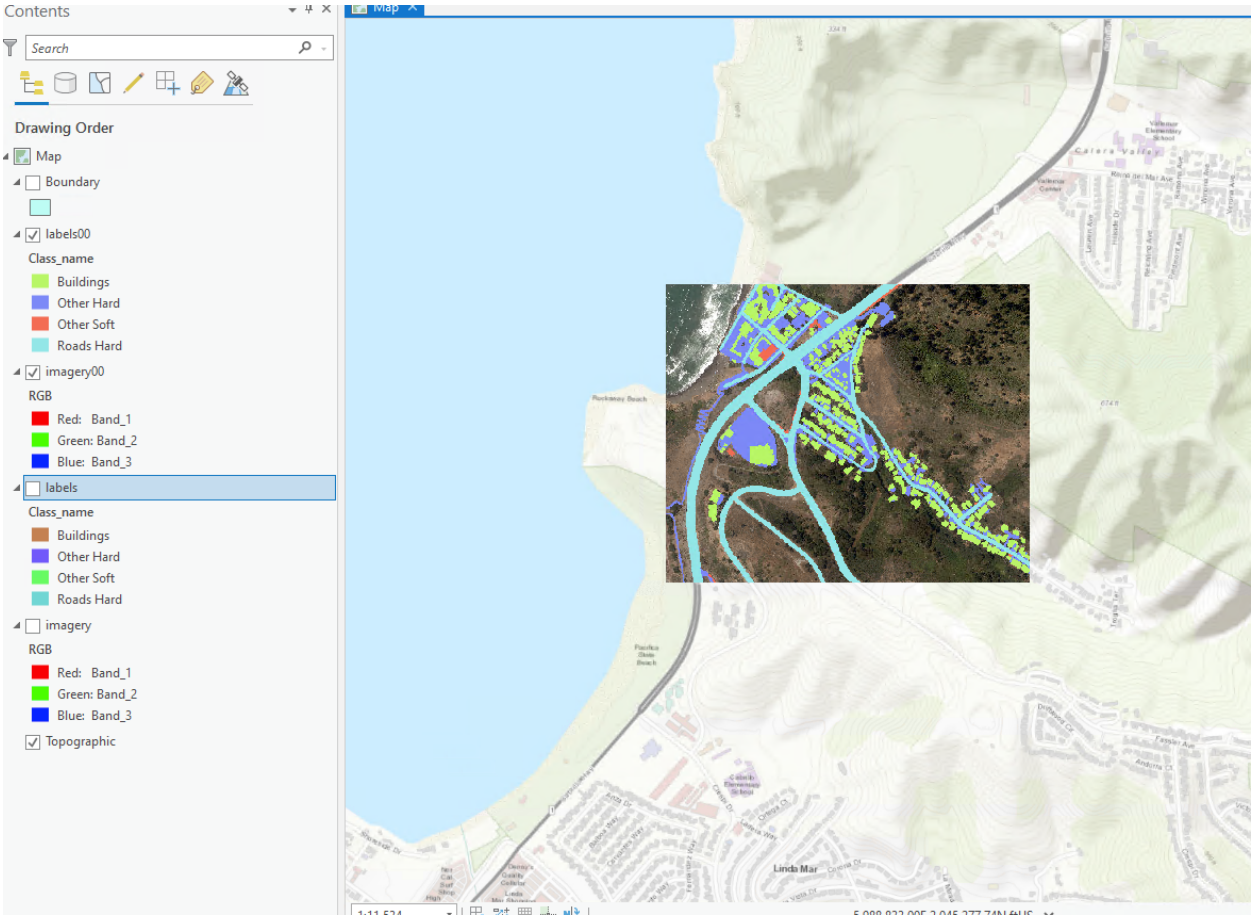
5. Find the Clip Raster tool and set up the parameters according to the screenshot below. Note: It is crucial to check the box to maintain clipping extent!



We will now do the same process with the labels, but this time we will open the environments menu on the Clip Raster tool and set the recently clipped imagery tile as the snap raster. Again, this ensures that the pixels on the imagery and labels match up perfectly. When clipping the rasters, it is common to run into 999999 errors, which can often be resolved by saving your project and edits, relaunching Arc, and running the tool again.

6. Repeat the process with the labels as described above.

At this point, you should have your imagery and labels clipped to a small section similar to the screenshot below.



It is important to check that both the cell size and the number of rows and columns of the imagery and labels are exactly equal. Right click on the imagery tile in the contents pane and navigate to Properties > Source > Raster Information. Do the same with the labels tile, and make sure the cell size and number of rows match each other. Even if they are only off by 1, it may cause problems when training your model.

7. Right click the imagery tile and hit export raster. Export the raster to a folder where you will store your training data, and repeat the process with your labels.

You now have your data prepared and ready to be sent to the VACC to train your model.