Master Thesis

Active Learning and Transfer Learning for the efficient Labelling and Semantic Segmentation in Aerial Imagery

Author:
Rajatha Nagaraja Rao

Matriculation Number:
223758

March 15, 2022

Advisers:
Konstantin Kirchheim  Leon Wehmeier  Prof. Dr. rer. nat. Frank Ortmeier

Department of Computer Science
Otto-von-Guericke University
Universitätsplatz 2
39106 Magdeburg, Germany
Rao, Rajatha Nagaraja:
Active Learning and Transfer Learning for the efficient Labelling and Semantic Segmentation in Aerial Imagery
Magdeburg, 2022.
Contents

Abstract

1 Introduction and Motivation
1.1 Motivation ............................................. 1
1.2 Aim of the thesis ...................................... 3
1.3 Structure of this thesis ............................... 4
1.4 Scope of this thesis .................................. 5

2 Background
2.1 Summary of Literature Review ....................... 7
2.2 Overview of Aerial-Datasets ......................... 8
   2.2.1 Open-Source Airborne Datasets for Semantic Segmentation 10
   2.2.2 Aerial Imagery and Orthophotographs ................. 13
   2.2.3 UAV and Satellite Imagery for Remote Sensing Applications 13
   2.2.4 Label Acquisition in Aerial Imagery .................. 15
   2.2.5 Analysis of Aerial Imagery and Orthophotos with Deep Learning ................. 15
2.3 U-Net for Semantic Segmentation in Aerial Imagery .......... 17
   2.3.1 Semantic segmentation in Aerial imagery using Deep Neural Networks .................. 18
   2.3.2 U-Net-based Architectures ....................... 19
2.4 Domain Shift with Transfer Learning .................. 22
   2.4.1 Advantages of employing Transfer Learning ........... 25
2.5 Active Learning for Efficient Data Labelling ........... 26
   2.5.1 Advantages of employing Active Learning ............. 29
2.6 Evaluation Metrics ................................. 30

3 Concept
3.1 Research Questions ................................. 31
3.2 Methods ............................................. 32
   3.2.1 Dataset Description and Domain Shift ............... 32
   3.2.2 Active Learning vs. Transfer Learning for Semantic Segmentation in Aerial Imagery ......... 34
   3.2.3 Proposed Model Architecture ..................... 39
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Prototype Implementation</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Dataset Preparation</td>
<td>41</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Data Labelling and Augmentation for the ObViewSly Dataset</td>
<td>41</td>
</tr>
<tr>
<td>4.1.2</td>
<td>Data Preparation for the Massachusetts Road and Building Datasets</td>
<td>42</td>
</tr>
<tr>
<td>4.1.3</td>
<td>Data Preparation Process for the LandCover.ai Dataset</td>
<td>42</td>
</tr>
<tr>
<td>4.2</td>
<td>Model Implementation</td>
<td>43</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Hyperparameter Tuning</td>
<td>43</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Technical Details</td>
<td>43</td>
</tr>
<tr>
<td>4.3</td>
<td>Model Training and Evaluation</td>
<td>44</td>
</tr>
<tr>
<td>4.4</td>
<td>Experiments</td>
<td>44</td>
</tr>
<tr>
<td>5</td>
<td>Evaluation and Discussion</td>
<td>47</td>
</tr>
<tr>
<td>5.1</td>
<td>Evaluation</td>
<td>47</td>
</tr>
<tr>
<td>5.2</td>
<td>Discussion</td>
<td>55</td>
</tr>
<tr>
<td>6</td>
<td>Conclusion and Future Work</td>
<td>59</td>
</tr>
<tr>
<td>6.1</td>
<td>Conclusion</td>
<td>59</td>
</tr>
<tr>
<td>6.2</td>
<td>Future Work</td>
<td>60</td>
</tr>
<tr>
<td>A</td>
<td>Detailed Results</td>
<td>61</td>
</tr>
<tr>
<td>A.1</td>
<td>Detailed results</td>
<td>61</td>
</tr>
<tr>
<td>A.2</td>
<td>Other preliminary experiments</td>
<td>65</td>
</tr>
<tr>
<td>B</td>
<td>Abbreviations and Notations</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>List of Figures</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>List of Tables</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>List of Algorithms</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>Bibliography</td>
<td></td>
</tr>
</tbody>
</table>
Abstract

Deep learning (DL) models are capable of performing semantic segmentation (SS) in aerial imagery that helps us detect the various semantic features such as buildings, roads, woodlands, water bodies and so on for use in several applications such as that of social and economic analysis. Since DL models are data hungry, data collection and ground truth generation are pivotal to training these DL models in a supervised setting. However, the annotations of interest may not be readily available and must be generated manually by annotating aerial images that span hundreds of kilometers which proves to be impractical on account of the extraordinarily high financial and temporal resources required.

This thesis work contributes towards minimizing the labelling effort by circumventing the manual labelling of the entire dataset through the use of both the transfer learning (TL) and active learning (AL) techniques. In the TL approach, an extensive research of available aerial datasets is done and DL models are pre-trained with suitable open-source aerial image datasets of varying ground sampling distances (GSD). The last layers of the models are then fine-tuned using a considerably small number of manually annotated samples in the ObViewSly dataset. The AL approach deals with label scarcity by iteratively selecting the most informative samples from an enormous unlabelled data pool which would help the model to learn better and to grow more confident in its predictions. We have used the entropy-based query strategy to rank and retrieve these samples for labelling which are then used for iteratively re-training the model. A novel technique called the AL guided TL (TL+AL) for SS in aerial imagery is proposed in this work that combines the effectiveness of both the AL and TL approaches. TL re-uses the learned representations from the source dataset and AL carefully selects important samples to be annotated such that we ensure both efficiency in labelling and good model performance. Also, the Shuffle-Unet model is proposed as a part of this work which employs phase shift in place of maxpooling and upsampling operations.

The AL, TL and TL+AL were investigated here with the U-Net and the proposed Shuffle-Unet models which were able to achieve an IoU score of 0.75 by judiciously annotating only 10% of the dataset. Through the entropy heatmaps, it was demonstrated that the samples that have regions covered by shadow are difficult to learn for the model. Variable GSDs lead to domain shift between the datasets used in TL. This domain shift is also addressed here wherein the trained model is adapted to detect semantic features in the ObViewSly dataset that has a lower GSD than that of the source dataset used for pre-training in TL. Hence, our approach attained good segmentation performance while incurring significantly low labelling costs.
Acknowledgements

I would like to express my sincere gratitude to my advisers Konstantin Kirch-heim, Leon Wehmeier and Prof. Dr. rer. nat. Frank Ortmeier for providing me with the opportunity to pursue my master thesis under their esteemed guidance. I am grateful for their mentorship and constant guidance which has helped me successfully undertake this endeavour.

I am immensely thankful to my parents and my friends for their constant encouragement throughout my thesis work and I earnestly thank the Otto-von-Guericke-Universität, Magdeburg for providing excellent education and outstanding support and opportunities to students.
Introduction and Motivation

This chapter presents an overview of the thesis and introduces the motivation that inspired this thesis work. The aim that would be accomplished through this work is elaborated and the structure and scope of the thesis work are outlined here.

1.1 Motivation

Images of the earth’s surface, captured with the help of satellites and unmanned aerial vehicles (UAVs), are utilised for remote sensing (RS) applications in several sectors such as defence, forestry, agriculture, weather monitoring, urban planning and so on. The wide applicability of these images across several domains makes them a vital source of information to learn about the earth itself. A noteworthy application of satellite/UAV images of land surface is that of socio-economic analysis which encompasses a variety of studies, with the aim of explaining how the economic undertakings and social development of a region influence each other (NOOR et al., 2018; SETO und KAUFMANN, 2003; YAO et al., 2019). Socio-economic data has different origins such as population census, economic census, land records, infrastructure records et cetera which can be obtained through in-situ and manual field surveys or through RS technology such as satellites and UAVs.

Monitoring of land use is one of the most important aspects of socio-economic analysis that helps to understand the elements of land development and sustainability, or lack thereof. The unprecedented change in

---

1 https://en.wikipedia.org/wiki/Socioeconomics
2 https://ibis.geog.ubc.ca/courses/klink/gis.notes/ncgia/ut08.html
land use patterns, especially in developing countries, has led to an increase in urban landscapes that are characterized by a reduction in woodlands, natural ecosystems and land area available for agricultural activities. This is a direct result of industrialisation and urbanisation wherein rapid increase in the utilisation of land for activities such as mining, housing and construction of other social infrastructure and industries can be observed. It is essential to monitor this rapid change in land use patterns and model the socio-economic impetus for this change so that land administrative policies for sustainable urban planning and development can be implemented (Hall und Wahab, 2021; Noor et al., 2018; Opalowa et al., 2014; Seto und Kaufmann, 2003). This thesis contributes to this endeavour by performing the automatic semantic segmentation (SS) in RS image data using deep learning (DL), with minimum overhead of image annotation, in order to identify various semantic features such as buildings, roads, woodlands et cetera.

Although there has been significant effort to automate the analysis of aerial imagery, yet a considerable amount of manual effort is currently required for the annotation of images pertaining to supervised machine/deep-learning (ML/DL) tasks. Creating image annotations is especially a mountainous task due to the high velocity, volume and diversity of the captured data. As satellites and UAVs are capable of observing vast expanses of land in a very short span of time, they can generate colossal volumes of data, especially in real-time that can be overwhelming for human-analysts to understand and gain valuable insights from. Hence, there is a special need for methods such as semi-supervised learning (SSL) that are capable of analysing high volumes of image data and deriving value from them at a quick pace with limited manual annotation efforts.

Due to the rapid development in computing hardware and large volumes of freely accessible data, the use of DL for computer vision (CV) is on the rise. DL provides better accuracy when compared to traditional CV techniques and helps to overshoot the tedious feature extraction process (O’Mahony et al., 2019). In order to perform CV tasks using DL algorithms, one requires large volumes of labelled data. As mentioned, the alarmingly high volumes of RS data make it a tedious and infeasible process to manually generate annotations for training DL models. The motivation of this thesis is to tap into the potential of utilising low volumes of labelled RS data
for the purpose of building CV-based deep-segmentation models using SSL techniques such as Active Learning (AL) along with Transfer Learning (TL).

1.2 Aim of the thesis

In order to harness the data present in imagery from RS platforms, and to utilise them for performing meaningful analysis with DL such as socio-economic analysis, there is an extensive need for a labelled dataset. Conducting regular land surveys helps to monitor current land use as well as to promote environmentally, socially and economically sustainable and effective land development and to enforce restrictions and regulations as necessary to prevent the illegal and unsustainable use of the available land area. Currently, such analysis can be performed more efficiently and accurately in terms of time and cost with the use of aerial imagery and orthophotos ie. RS data coupled with computer vision-based deep learning (CVDL) algorithms.

Labelled datasets are difficult to generate due to the required tedious manual labour, and the incurred time and monetary costs, especially for the task of SS, because it requires pixel-level annotations. For instance, an image of 1000x1000 pixels requires 1 million pixel annotations to make it suitable to be used for a supervised segmentation task. Although difficult to generate, it is immensely important to produce these annotations as they are essential to ‘understand’ the different land-objects such as roads, buildings, woodlands, and vehicles present in the RS image data and to train deep-segmentation models. Therefore, it is comparatively more feasible to annotate a small, initial dataset which can further be carefully utilised in DL models using techniques that are capable of working with small amounts of data. It would be an extraordinary endeavour to manually generate pixel-level annotations for semantic features present in images spanning hundreds of square kilometers. Therefore, the objective is to circumvent this tedious manual labelling process by employing DL models to automatically identify the different ground features with the help of similar, pre-existing, labelled datasets and through the careful selection of important image samples to annotate, that would result in better learning by the DL model.

In order to minimise the labelling effort and to ensure a good perfor-
mance of the DL models, AL and TL approaches for the task of SS are implemented and their performances are then compared for the given problem. The aim of this thesis is to identify different semantic features present in RS image data from the ObViewSly dataset through SS with DL, and hence augment to the existing feature-set through techniques like AL and TL that work well with substantially small datasets.

Currently, there is a lack of extensive work in the field of comparing and combining AL and TL methods (TL+AL) for SS, especially for aerial imagery. Hence, as a part of this thesis work, we propose a novel approach wherein different experiments are performed individually and by combining the two techniques to teach the TL model through AL, and evaluating their results on the ObViewSly dataset.

1.3 Structure of this thesis

This section outlines the structure of the thesis work which consists of 6 chapters.

Chapter 2 discusses the fundamental concepts that serve as the basis for this thesis work. It provides an overview of aerial imagery, deep learning architectures, learning techniques and their applications in the analysis of aerial imagery. A review of previous research undertaken in this field is also mentioned here.

Chapter 3 describes the research questions addressed in this thesis work and formally describes the datasets. The methodological approach to address these research questions such as the proposed algorithm and proposed model architecture are described in detail.

Chapter 4 provides the technical details for implementing the methods proposed in chapter 3 through experiments, so as to address the research questions defined. It includes the details of dataset preparation, hyperparameter tuning, model training and evaluation, and information regarding deep learning libraries used and the underlying computational resources.

Chapter 5 presents the detailed results and the evaluation of the experiments conducted in this work.

Chapter 6 presents the concluding remarks for the overall thesis work.
Several possibilities for further improvements and applications are provided as a part of the future outlook.

1.4 Scope of this thesis

This thesis work encompasses several efforts towards CVDL for SS in aerial imagery. In particular, the focus is to perform multi-class SS by training DL models to identify 4 semantic classes namely, road, building, green-cover and background. Research, implementation, evaluation and comparison of the different approaches in SS in aerial imagery with a limited-size labelled dataset, through AL and TL techniques is performed. This work is geared towards performing a comparison of AL and TL techniques with U-Net based architectures, with the objective of augmenting to the existing 'building' label in the ObViewSly dataset, such that various semantic features can be accurately identified in the ObViewSly dataset without the need to exhaustively create annotations for different objects in the whole dataset.

For the TL approach, the goal is to research and identify annotated RS datasets pertaining to aerial imagery with top/oblique views in urban/suburban areas, that can be used to pre-train custom deep-segmentation models that would be capable of identifying semantic features in the ObViewSly dataset after fine-tuning through TL techniques. Pixel-level annotations for selected images in the ObViewSly dataset are manually generated, to facilitate the process of fine-tuning the custom TL models and to iteratively provide informative, labelled samples in the AL pipeline. The AL pipeline is implemented according to the available time and computational resources to support iterative querying and model re-training. The models are then evaluated based on mean Intersection over Union (mIoU) metric on every iteration of the different approaches such as AL, TL, TL+AL for the purpose of comparing their performances on the ObViewSly dataset.
The Background chapter is divided into sub-sections to elucidate the formulation of the approach followed in this thesis and to explain some of the related work carried out in the field of deep learning for semantic segmentation with remote sensing image data. The literature review and the theoretical background pertaining to this field of research is presented here.

- Summary of Literature Review
- Overview of Aerial-Datasets
- U-Net for Semantic Segmentation in Aerial Imagery
- Domain Shift with Transfer Learning
- Active Learning for Efficient Data Labelling
- Evaluation Metrics

### 2.1 Summary of Literature Review

This section presents the literature review process and the theoretical concepts in the background work that serve as a basis for designing and conducting the different experiments in this thesis. In all of the cases, the most relevant and latest papers (according to subject-wise availability) were chosen and snowballing and reverse snowballing techniques were employed to analyse and choose related works as necessary. Further, this section presents the studies that are most applicable to this thesis in a tabular format. These studies are further explored in the upcoming sections.
Table 2.1: Survey of aerial datasets and deep learning applications

<table>
<thead>
<tr>
<th>Paper</th>
<th>Aim of the study</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>(MOHANTY et al., 2020; SHETTY et al., 2020; ULMAS und LIJV, 2020)</td>
<td>Performs DL-based classification and SS in airborne imagery in order to uncover the different features on the land surface with the goal of understanding socio-economic status of a region.</td>
<td>BigEarthNet; SpaceNet Satellite Imagery dataset; Custom dataset created with MapPuzzle software</td>
</tr>
<tr>
<td>(PERSELLO und KUFFER, 2020)</td>
<td>Maps a given image to a socio-economic index (DIMD) using DL.</td>
<td>OpenStreetMap, VIIRS, WorldPop and other health and demographic census data.</td>
</tr>
<tr>
<td>(MNIH, 2013)</td>
<td>Experiment with traditional and DL-based CV algorithms and to curate an urban, aerial dataset using OSM</td>
<td>Massachusetts Road and Building datasets</td>
</tr>
<tr>
<td>(AZIMI et al., 2019)</td>
<td>Curation of a dataset of highly granular annotations and comparison of the performance of several state-of-the-art segmentation models</td>
<td>SkyScapes</td>
</tr>
<tr>
<td>(BOGUSZEWSKI et al., 2021)</td>
<td>Generation of a dataset with granular, pixel-level annotations in RGB aerial images</td>
<td>LandCover.ai</td>
</tr>
<tr>
<td>(WANG et al., 2016)</td>
<td>Capture images from several views of the city to produce a benchmark dataset.</td>
<td>TorontoCity Benchmark dataset</td>
</tr>
<tr>
<td>(ADIBA et al., 2019; HE et al., 2019; WU et al., 2019)</td>
<td>Performs different transfer learning techniques in U-Net-based architectures for the purpose of SS in images.</td>
<td>Assortment of aerial imagery that are publicly available and/or those generated by the authors for specific usecases.</td>
</tr>
</tbody>
</table>

2.2 Overview of Aerial-Datasets

There has been immense progress in airborne photography since its inception as pigeon photography in 1907 by Julius Neubronne. Aerial imagery or airborne imagery acquires the top/oblique bird’s-eye view of land features from RS platforms that are airborne such as UAVs, piloted aircrafts.
and satellites. High-resolution images, when combined with economic and demographic data for the region and analysed together can generate further insights beyond those obtained solely with RS data (Seto und Kaufmann, 2003). Frequent surveying of land assists in the maintaining of an up-to-date database of land information and monitoring the transformation of land Year-Over-Year (YOY) and in identifying and performing econometric studies on the driving forces causing these changes in land use (Opaluwa et al., 2014). As one of the goals of the thesis is to augment to the existing label-set in our ObViewSly dataset, by circumventing the process of large-scale manual image annotation, as it is extremely time consuming and requires significant manual efforts and pixel-level accuracy in the annotations, it is essential to identify the existing sources of data with accurate annotations, and to investigate different learning techniques that can work efficiently with small amounts of target domain data so as to reduce the labelling effort. In this regard, the following investigations were made in identifying suitable datasets for the task.

The unprecedented change in land use patterns in recent times, especially in developing countries, has led to a transformation in the view of the earth’s surface leading to increased urban landscapes (Opaluwa et al., 2014; Seto und Kaufmann, 2003). Land use patterns can be rapidly monitored in real-time with data such as aerial imagery from RS platforms. This provides a faster alternative to in-situ measurements.

A plethora of datasets are freely available on different platforms and the drawback observed is that although there are several generic datasets available such as ImageNet, SpaceNet, BigEarthNet datasets, MS-COCO etc, there is still a scarcity in the availability of aerial images with top-down view containing task-specific annotations and they are also tedious to obtain. We investigate the utility of different aerial datasets, based on the quality of the available data, along with the availability and the quality of annotations for the purpose of SS. Different datasets provide different annotations and these have been thoroughly analysed and considered accordingly for utilisation in our purpose of image interpretation in the ObViewSly dataset.

Spatial resolution of the aerial image is measured in terms of Ground Sampling Distance (GSD) wherein GSD represents the distance between image pixel centers as measured on the ground and hence helps to understand the scale of the semantic objects on the ground. Thus, a GSD of 1
meter indicates that each pixel of the image depicts an area of 1 square meter on the ground. Changes in GSD results in varying granularity of the semantic features present in the image. Lower GSD values indicate a higher spatial resolution in the aerial images and hence capture semantic features at a higher granularity with higher level of detail.

There is an ever-growing need for annotated datasets that would facilitate the training of ML/DL models to be employed in several different tasks. During the process of obtaining suitable datasets with an acceptable GSD that corresponds with that of the ObViewSly dataset, it was observed that there is a lack in the availability of assorted datasets that can be employed in SS tasks. The following survey was compiled during the research for a labelled dataset that has comparable features as that of the ObViewSly dataset in terms of GSD and semantic features such as those in urban/suburban/rural areas, along with the necessary pixel annotations for learning the semantic features depicted in the images.

### 2.2.1 Open-Source Airborne Datasets for Semantic Segmentation

The dataset under consideration, ObViewSly, provides top-down view aerial images of suburban and urban areas in Germany. It has a low GSD of 10cm which depicts the ground objects with high granularity. In order to augment labels to this dataset through different techniques such as TL, there is a need to assemble similar datasets for the purpose of obtaining semantic labels. During this research, it was observed that there is an increase in the availability of datasets that facilitate the study of SS for autonomous driving. These datasets do not provide a top-down aerial view but rather provide a front-view of lanes such as the SYNTHIA dataset (Ros et al., 2016). Only datasets that provide top-down and oblique views were considered in this work as described below:

- **Massachusetts Road and Building datasets** - Several efforts have been invested towards the subject of Machine Learning for aerial image labelling. This includes the work of (MNHH, 2013) wherein different ML/DL techniques are discussed for the purpose of aerial image labelling and DL-based segmentation and two new GeoTiff datasets known as Massachusetts Buildings Dataset with 151 aerial images and Massachusetts Roads Dataset with 1171 aerial images spanning
2.2. OVERVIEW OF AERIAL-DATASETS

(a) Input RGB image

(b) Combined road and building annotations

(c) Building annotations

(d) Road annotations

Figure 2.1: Massachusetts Road and Building datasets

urban, suburban and rural areas, each of size 1500x1500 are curated with the help of Open Street Maps (OSM). The building footprints and road centerlines were rasterized to obtain the pixel-annotations. This dataset, as seen in figure 2.1 has a ground sampling distance of 1 meter ie. each pixel covers one square meter and it has been utilised in experiments for addressing the first research question of domain shift in the thesis.

• Skyscapes dataset -(AZIMI et al., 2019) introduces the Skyscapes dataset which provides highly accurate and highly granular pixel annotations for 31 different categories of objects. This dataset is also
evaluated against state-of-the-art deep-segmentation models and a novel DL model architecture is proposed to handle datasets with a varied range of scales, objects and levels of granularity. Although the dataset is commendable, enormous amounts of time of approximately 200 man-hours were consumed for the annotation of every image. Also, the dataset is available for access only after obtaining prior permission on request.

- **LandCover.ai dataset** *(Boguszewski et al., 2021)* proposes the publicly available LandCover.ai aerial imagery dataset which provides highly accurate manual pixel-annotations for 5 classes namely buildings, roads, water, woodlands and background for the corresponding RGB GeoTiff images as seen in figure 2.2, using the EPSG:2180 spatial reference system, of regions in Poland. This orthophoto-dataset fills the existing need for semantic labels covering suburban and rural areas. Each GeoTiff tile is available at a resolution of 9000x9500 px or 4200x4700 px. This dataset consists of highly granular representations of objects as the images are acquired with a GSD of 25cm and 50cm, and thus serves as a good starting point to train TL models to identify features on the ObViewSly dataset. This dataset was conducive to address the first research question of domain shift in the thesis work.
2.2.2 Aerial Imagery and Orthophotographs

Aerial Imagery and Orthophotography generate images representing top and/or oblique views using RS platforms such as UAVs and satellites (Noor et al., 2018). This has been made possible due to technological advancements in the miniaturisation of necessary instruments to capture these photographs. Aerial images innately possess certain disadvantageous characteristics such as displacements and distortions due to camera lens or camera tilting, and do not reflect true ground distances ie. it does not embody a uniform scale. Hence, these negative properties are corrected during the generation of orthophotos through geometric transformations and are scaled to reflect true measurements such as angels and distances (Rabiu und Waziri, 2014). Hence, true distances can be measured using the Engineer's Scale on the orthophoto\(^1\). Also, orthophotos create a photographic map with a straight-down view of a given area as they are projected onto the earth's coordinates using various Coordinate Reference Systems (CRS) formats such as Universal Transverse Mercator (UTM), International Mapcode system and so on. This process is also known as orthorectification and it is worth mentioning that this orthorectification process of projecting aerial images to create the map of a region requires significant effort (Mohanty et al., 2020).

Hence, aerial images and orthophotos can be manually or automatically analysed to understand the objects on the ground such as various land features and their distributions. This in turn can facilitate socio-economic analysis such as change in land cover and setting up of necessary policies accordingly in order to prevent the depletion of natural resources, unsustainable infrastructure development and other issues of impractical land use.

2.2.3 UAV and Satellite Imagery for Remote Sensing Applications

Cities and countries span enormous areas of land and certain areas of land are inaccessible due to various factors such as rough terrains, remote locations and military areas. Here, it is inconvenient to perform in-situ measurements and surveys in some cases. Also, with remote measurements

\(^1\) https://arizonasurveying.com/orthophotography/
with UAVs (third-generation RS platforms) and satellites (second-generation RS platforms), a larger area of land can be viewed at once. This helps to generate both highly granular measurements in case of high-resolution imagery, as well as to generate an overview of the land area under scrutiny. It also creates a platform for unbiased data collection (NOOR et al., 2018).

Although satellite imagery is especially useful in tracking global human footprint through Meteorology, Forestry, Cartography to name a few, UAVs have carved a niche for themselves as a ubiquitous necessity. UAVs are capable of obtaining images from low altitudes at very high resolutions and can be adapted to different needs and locations - indoors, outdoors, underwater and provide different options of flight ranges whereas these features are hard to match by satellite-based imagery. Undeniably, there exist several further uses of UAVs for varied purposes such as digital surface modelling (DSM), Photogrammetry and so on for leisure, military, academic and other purposes (HALL und WAHAB, 2021) which are outside the range of this thesis. Although satellites with the capacity of ultra-high spatial resolutions are deployed, they are not as affordable and spatio-temporal-spectrally flexible as deploying drones for similar motives, with safety and ethical deliberations for RS.

Spatial resolution is an important factor in aerial imagery and orthophotography as it informs us about the area on the ground represented by a single pixel in the image also known as GSD. Currently, it is possible to achieve a very high spatial resolution as granular as 1cm GSD and to acquire ultra-high resolution imagery (NOOR et al., 2018). Due to the miniaturization and availability of a wide range of components such as cameras and sensors to capture data with UAVs, with spatio-spectral-temporal adaptability, it is possible to custom-make these RS platforms to fit the required niche. This helps in easier and more efficient data acquisition which enables to acquire finer details of the objects on the ground faster.

Some drawbacks are still observed. Due to the high volume of images, it is not feasible to manually label every image and pixel for the purpose of understanding the acquired data. Due to the atmospheric effects in urban areas, and inaccuracies in capturing and mapping the data, there is still a requirement for appropriate field verification in some cases.
2.2. OVERVIEW OF AERIAL-DATASETS

2.2.4 Label Acquisition in Aerial Imagery

There are several possibilities for acquiring semantic labels for aerial images such as those available through public registries and governmental agencies. The most common but severely taxing method is to perform manual annotations according to the task-based requirements. As this incurs significant costs and also extends the completion time of projects, we turn to other methods through which these annotations can be acquired. As DL has become the de-facto standard for SS and deep neural networks are data-hungry, a large number of training examples must be provided in order to build models with high performance. Hence, the label acquisition process becomes extremely important as models can only improve their performance when good quality and sufficient data and ground truth are provided.

Weakly supervised ground truth can be provided through incomplete supervision, inexact supervision and inaccurate supervision, when it is cumbersome to generate strong supervision data such as complete annotations (ZHOU, 2018). Several sources that help to generate ground truth for aerial imagery include OSM and Google maps. (KAISER et al., 2017) take advantage of the large, freely available data from OSM and Google Maps to generate weak annotations. This is based on the hypothesis that large volumes of probably noisy data with a wide range of samples can improve the generalisation capability of the model and improve the classification performance.

In this thesis, manual effort is applied in producing pixel annotations for selected images in the ObViewSly tiff files for the purpose of training deep-segmentation models in a TL and AL setting.

2.2.5 Analysis of Aerial Imagery and Orthophotos with Deep Learning

A special interconnect exists among the various aspects of airborne image data acquisition, its analysis with ML/DL techniques and their culmination of socio-economic research. As it was previously established, it is easier and cost-effective to capture numerous, high quality images with UAVs at low GSD or high spatial resolution when compared to coarser, low spatial resolution image data obtained from satellites. Highly accurate
representations of different land features can be obtained and this in turn helps to efficiently identify the different objects present in the image such as lanes, buildings, forest cover et cetera which helps to predict the socio-economic standing of a region. This task can be carried out automatically with ML/DL.

Application of DL-based models - especially that of Convolutional Neural Networks (CNNs) have proven to be effective in automatically understanding and extracting insights from image data after being trained on an initial image dataset ie. CVDL tasks. These applications include image classification, SS, object recognition and instance segmentation tasks. Application of CNN architectures such as U-Net and Mask R-CNN, along with architectural modifications such as Conditional Random Fields (CRF), filters have proven to automate aerial image annotation (Mohanty et al., 2020).

Neoteric ideas of utilising DL for socio-economic analysis using airborne imagery have been on the rise and some of the studies are presented here. All of the following studies employ DL algorithms on airborne imagery obtained from Satellites/UAVs.

1. (Shetty et al., 2020) propose the segmentation of land features through DL and traditional image processing techniques such as pixel-wise thresholding. Here, multiple pre-processing steps were performed on the RGB image such as canny edge detection, bilateral filtering, grey-scale conversions et cetera and their results are combined with that of CNN model results to predict the socio-economic status. The CNN model proposed in this paper achieved an accuracy of 90% in detecting buildings, water, agriculture and roads.

2. (Persello und Kuffer, 2020) uncover the level of socio-economic inequalities of deprived neighbourhoods/slums that lack access to basic infrastructure such as sanitation and housing, without the need for extensive ground surveys by adapting the VGG-16 model trained for slum classification, through TL. The deprivation index of the regions is calculated based on the Data-driven Index of Multiple Deprivation (DIMD) framework (Ajami et al., 2019) and this index can be extrapolated over other unseen urban areas.
3. (ULMAS und LIIV, 2020) acknowledged the bottleneck of inavailability of high volumes of labelled training data and the difficulty in obtaining pixel-level training masks. This paper claims that the model learns better if it is initially trained with relevant data such as that of BigEarthNet’s Sentinel-2 data as opposed to generic datasets such as ImageNet. They leveraged the BigEarthNet classification dataset for training a ResNet50 classification model, that achieved a high $F_1$ score, and utilised it as a pre-trained encoder in U-Net for segmentation, through TL. This segmentation model made use of a small, custom dataset curated using Sentinel-2 images and CORINE land cover map for model fine-tuning, and showed a high IoU performance on the test data.

4. According to (LEVY et al., 2021), life expectancy can be predicted based on features such as income, education, age et cetera of neighbourhoods using CNNs. Here, it was seen that TL techniques based on ResNet50 architecture, initially trained on the publicly available ImageNet dataset, to detect different kinds of objects could be re-trained to predict mortality rates per image. These predicted mortality rates were averaged across the county and compared with the true county-level mortality rates for model training. This paper proves that, other sources of demographic data can be incorporated with image data for use in applications pertaining to similar use cases.

This thesis aims, as mentioned in the above use cases, to perform a similar examination of the airborne image datasets. Specifically, the aim is to enrich the collection of feature annotations for the ObViewSly dataset. This thesis performs TL and AL-based experiments in order to identify and append to the ‘building’ annotations, features such as roads and green cover by utilising DL models.

2.3 U-Net for Semantic Segmentation in Aerial Imagery

This section presents the basic concepts for semantic segmentation and its use in aerial imagery. Here, the suitable DL network architectures are also discussed.
2.3.1 Semantic segmentation in Aerial imagery using Deep Neural Networks

Image segmentation is one of the most challenging tasks in CV (Guo et al., 2018) and plays a vital role in CV and image processing. The process of SS can be viewed as a pixel-level classification task wherein every individual pixel in the image is assigned a separate label based on the class that it belongs to. The superiority in robustness and accuracy of DL methods, especially that of convolutional neural networks (CNNs) for SS make them the most opted technique wherein the input layer is followed by several hidden layers with a final output layer which outputs a number of channels which is equal to the number of necessary pixel classes (Goceri, 2019).

DL makes it easier to work with image-processing tasks as it performs the automatic feature extraction process and learns the representation for the data. Hence, there is no necessity for the separate extraction of features before ingesting images into the DL model. Also, the use of CNNs drastically reduces the number of parameters learned by the model due to parameter sharing (Goodfellow et al., 2016) of filter weights that are applied across the entire image. This also speeds up the training process while reducing the memory consumption during model training. Unfortunately, one major drawback is constantly observed which inhibits harnessing the capabilities of these DL algorithms in CV tasks. It is that of lack of sufficient amount of task-specific labelled datasets for model application (O’Mahony et al., 2019). Hence, it is the need of the hour to build models using techniques that have the capacity to build powerful models with limited amounts of data available. Another drawback that is observed is that a model trained for one task is unable to generalise to a task in a different domain. We aim to address these two drawbacks in this work by addressing domain shift and generalisation in DL models as well as to explore techniques that can be useful to build models with good performance when there is a limited dataset for SS.

Several DL models are proposed that achieve high levels of performance across several metrics such mIoU and Area Under the ROC curve (AUC) for the task of SS. Currently, several state-of-the-art models exist for performing SS and these include DeepLabV3+, SegNet, W-net, U-net, V-net, and ResNet, VGGnet based architectures and ensemble models such as those that utilise
different encoder-decoder network combinations such as (ResNet, VGGnet) (Badrinarayanan et al., 2017; Chen et al., 2018; Lateef und Ruichek, 2019; Ulku und Akagunduz, 2019). In this work, we consider the U-Net architecture along with necessary modifications to compare the different approaches of TL and AL for the goal of SS and hence the corresponding label augmentation in the ObViewSly dataset in order to understand the different semantic features present in the image for further analysis.

2.3.2 U-Net-based Architectures

U-Net

The U-Net, as depicted in figure 2.3 is a famous encoder-decoder architecture that was proposed for the task of SS in medical images, is a fully convolutional network (FCN) made up of a contracting and expanding path connected by skip connections (Ronneberger et al., 2015). U-Net-based DL models generally consist of a U-shaped path wherein the input image is initially downsampled while doubling the number of convolutional, 3x3 filters in order to learn the representations accurately. The data is then once again upscaled in order to obtain the pixel-wise segmentation masks in the output, corresponding to the same width and height dimensions as that of the input image. The contracting path (downsampling path) reduces the spatial dimensions with the help of max pooling layers. In order to ensure that the low level features are propagated further along in the network, skip connections are utilised, wherein the feature maps from the contracting path are concatenated in a channel-wise manner with the corresponding, same level feature maps of the expanding path. U-Net-based architectures work very well for SS in several applications such as medical images - MRI, Ultrasound data, aerial images from RS platforms - satellites, UAVs, and several other applications that focus on segmentation tasks. One observed drawback is the presence of checkerboard artifacts introduced due to the deconvolution layer used for upsampling in the expansion path ie. due to the deconvolution overlap that occurs when the kernel size is not divisible by the stride or due to random initialization (Aitken et al., 2017). Although pooling layers help to achieve translation invariance, information loss is observed due to the pooling
operation such as max pooling which considers only the higher value pixels.

Figure 2.3: U-Net Architecture (RONNEBERGER et al., 2015)

The periodic shuffling operation

During the process of upscaling an image from low resolution (LR) to high resolution (HR) space, there exists a problem of the introduction of checkerboard patterns into the HR image, due to the deconvolution overlap. Also, the introduction of 0 values while performing upsampling with strided convolutions means that these values must be further improved as the network learns through backpropagation. Using the periodic shuffling operation for upsampling after the convolution process helps to overcome the checkerboard patterns that appear in the output image (LACHINOV, 2019). The introduction of 0 values can be avoided by utilizing several filters followed by resizing of the activation map to produce an upscaled image. Therefore, the use of phase shift/periodic shuffling operation for performing the upscaling in the expanding path of the U-Net for the SS in aerial imagery is proposed in this thesis work. The periodic shuffling operation is a phase shift/image reshaping operation only that is used to
rearrange a $H \times W \times C \cdot r^2$ tensor to a $rH \times rW \times C$ tensor where $H =$ height of the image, $W =$ width of the image and $C =$ number of channels and $r =$ upscaling factor. This operation is formally defined as (Shi et al., 2016):

$$PS(T)_{x,y,c} = T_{\lfloor x/r\rfloor,\lfloor y/r\rfloor, C \cdot r \cdot \text{mod}(y,r) + C \cdot \text{mod}(x,r) + c}$$ (2.1)

where $x,y$ are the output pixel coordinates in the HR space, $c$ is the number of channels and $r$ is the scaling factor.

The periodic shuffle/pixel shuffle\(^2\) takes an input of $(C_{in}, H_{in}, W_{in})$ and generates an output $(C_{out}, H_{out}, W_{out})$ where $C_{out} = C_{in} \div r^2$, $H_{out} = H_{in} \times r$ and $W_{out} = W_{in} \times r$ where $r$ indicates the scaling factor, $C$ indicates the number of channels in the tensor, $H$ and $W$ represent the height and width of the tensor respectively the the $x,y$ coordinates within each spatial block of the output as indicated by figure 2.4 are determined by the high order component of the input channel index. Hence, it rearranges the pixels and transforms the feature maps from a LR domain to a HR domain without involving convolution operation and this operation is not random (Qin et al., 2020). The pixel unshuffle operation reverses the shuffle operation wherein it takes an input of $(C_{in}, H_{in}, W_{in})$ and generates an output $(C_{out}, H_{out}, W_{out})$ where $C_{out} = C_{in} \times r^2$, $H_{out} = H_{in} \div r$ and $W_{out} = W_{in} \div r$ and maps the feature maps from the HR to LR space without loss of pixel values, just by rearranging it by dividing the input channel into patches and concatenating them along the channel dimension. Also, the use of pixel unshuffle operation in place of maxpooling operations is free-from information loss caused due to the maxpooling operation as the pixels are

only reordered (Chatterjee et al., 2021). It is to be noted that the shuffle and unshuffle operations are not learnable.

### 2.4 Domain Shift with Transfer Learning

In general, one can observe a degradation in the performance of a DL model when there is a shift in the distribution of data. Domain shift is generally observed when there is a change in the distribution between the training data (data that is used to train the DL model) and the test data that is encountered when the model is either deployed or the model is subjected to as a result of a general change in the distribution of the dataset observed over time. Here, domain shift refers to the difference in the data distribution and feature space in the source and target domain datasets. Specifically, the domain shift arises when we train deep-segmentation models on aerial imagery that spans different landscapes such as that of urban or suburban and rural areas across different ground sampling distances. Aerial images acquired at 10cm GSD possess drastically different features from those acquired at a GSD of 100cm. Hence, a learner trained on one of the domains may not possess good predictive capabilities on the other. TL helps to bridge this gap by facilitating the learner to adapt and incorporate a new representation of target domain data, especially in cases wherein there is limited availability of data. A general assumption while training supervised ML/DL models is the availability of a large dataset, with accurate ground truth information. Sometimes, it can be challenging to obtain sufficient data due to several hindering factors that prevents us from training task-specific models directly (Yao et al., 2019). In these scenarios, a practical approach is to train models on an existing dataset with available annotations and to further adapt these models to perform domain-specific tasks.

Here, TL is formally defined as (Pan und Yang, 2009):

*Given a source domain $D_s$ and a source task $T_S$ and a target domain $D_t$ with the target task $T_t$, TL is the process of improving the target predictive function $f_T(\cdot)$ in $D_T$ by using the knowledge from $D_S$ and $T_S$, where $D_s \neq D_t$ or $T_s \neq T_t$.*
Algorithm 1 Domain Adaptation with Transfer Learning (TL)

Given:
\[ x_{S_i}, y_{S_i} \text{ where } x_{S_i} \in X_S \text{ and } y_{S_i} \in Y_S \text{ in } D_S = \{(x_{S_1}, y_{S_1}), ..., (x_{S_n}, y_{S_n})\} \]
\[ x_{T_i}, y_{T_i} \text{ where } x_{T_i} \in X_T \text{ and } y_{T_i} \in Y_T \text{ in } D_T = \{(x_{T_1}, y_{T_1}), ..., (x_{T_n}, y_{T_n})\} \]
such that \( 0 < n_T \ll n_S \)

Model Pre-training:
Learn a model \( f_S(\cdot, \theta) : X_S \rightarrow Y_S \) with parameters \( \theta \)

Model Fine-tuning:
Fine-tune the parameters \( \theta \) in \( f_S(\cdot, \theta) : X_S \rightarrow Y_S \) to learn a model \( f_T(\cdot, \theta_T) : X_T \rightarrow Y_T \) with fine-tuned parameters \( \theta_T \)

A source domain is a pair \( D = (X, P(X)) \), and \( Y \) is the label space such that the training data consists of pairs \( x_i, y_i \) where \( x_i \in X \) and \( y_i \in Y \) and \( f(\cdot) \) is the objective function such that \( f(x) = P(y/x) \). As seen in 1, \( x_{S_i} \in X_S \) is a data instance and \( y_{S_i} \in Y_S \) is its corresponding label in the source domain \( D_S = \{(x_{S_1}, y_{S_1}), ..., (x_{S_n}, y_{S_n})\} \). Similarly, the target domain \( D_T = \{(x_{T_1}, y_{T_1}), ..., (x_{T_n}, y_{T_n})\} \) contains \( x_{T_i} \in X_T \) and its corresponding labels \( y_{T_i} \in Y_T \). Mostly, \( 0 < n_T \ll n_S \) indicating that the number of available annotated data samples in the source domain is much larger than that in the target domain.

TL techniques are those in which a model is trained on an available dataset from the source domain that may or may not be similar to the data pertaining to the target domain, and this model is further fine-tuned or used directly to make predictions on data from the target domain. Hence, knowledge learned from the source domain is applied on the target task. This technique is highly useful in situations where a small amount of task-specific target domain data is available and it is expensive to generate further labels. In these cases, the pre-trained models can be fine-tuned on a small, target dataset to improve its prediction accuracy in the target domain.

(WEISS et al., 2016) states that TL can be used to improve the performance and generalization of a learner by making it applicable in a different domain than the one in which it was trained. This works due to the common characteristics that exist between the source and target do-
mains. Pre-trained models can be used as feature extractors that can learn the representation of the dataset and only a few of the layers may be further finetuned to perform the domain-specific task. This requires that the source and domain data have similar characteristics. The final layers of the model can be re-trained in order to learn the target dataset while retaining the representations learned by the initial layers of the network that learn the low level features such as edges present in the given image dataset.

Domain adaptation can be achieved using TL by re-using the representations learned on the source dataset and fine-tuning the model to adapt in order to learn the representations in the target input-output pair. The type of domain adaptation in SS performed in this thesis is also known as Closed Set domain adaptation as all the possible categories for pixel classification appear in both the source and the target domain datasets (Toldo et al., 2020) ie. roads, buildings, green cover and background.

The following provides the overview of use cases in which TL techniques are used for SS to address domain shift:

• In (Wu et al., 2019), a dataset of aerial imagery with a GSD of 0.5m was developed along with manually labelling 256x256 patches of image tiles. It applies TL techniques for U-Net and FCN models pre-trained on a similar dataset and compares their performances in terms of pixel accuracy and mIoU. It was observed here that U-Net identifies richer semantic details and was adapts to a different domain by retaining the representations learnt by deeper layers of the U-Net and fine-tuning the initial and final layers of the model. Similarly, (He et al., 2019) address TL in cross-modality image data in different spectra wherein the initial layers of the U-Net are fine-tuned and the representations learned by the deeper layers are retained. Further, TL in U-Net is employed for building segmentation in aerial imagery (Adiba et al., 2019). This paper iteratively fine-tunes different layers starting with the last layer of U-Net and increasing the number of unfrozen layers by 1 for every epoch until convergence, instead of fine-tuning all the layers at once to avoid catastrophic forgetting.

• Extending the usecase to biomedical image segmentation with DL, extensive analysis is performed in the fine-tuning of U-Net for image segmentation in ultrasound images (Amiri et al., 2020). Although it is
common practice in TL to retain the weights learned by the shallow layers of the deep-segmentation model and to fine-tune the deeper layers that are capable of identifying high-level image features, it was observed here that fine-tuning the contracting path i.e. the shallow layers of the U-Net was more beneficial to adapt to domain shift than training the deeper layers.

These papers serve as a basis for the TL experiments undertaken in this thesis work.

2.4.1 Advantages of employing Transfer Learning

• TL helps to utilize the knowledge learned from the source domain data and the source task in order to improve the learner on the target task in a related target domain and in turn helps to address the problem of domain adaptation.

• TL techniques help in applications where certain ground truth ‘X’ is available only in the source domain and we would like to train a learner to identify these labels in a target domain so as to augment to its existing feature-set, without the need to create labels ‘X’ in the target domain. This is a benefit as the time/monetary costs for generating annotations can be conserved. This is also true in the case of labelling aerial imagery as they span vast areas and significant man power is required to annotate the images for training a learner. TL is especially useful in use cases pertaining to domains wherein specialized skill and domain expertise is required for manual image annotations and acquiring labels can be very expensive and time consuming.

• TL models serve as a good initialization point for network weights that helps the learner to improve faster to adapt to the target domain. Also, several pre-trained models are available and it requires lesser data and computing power to fine-tune them Hu1 (2020).
2.5 Active Learning for Efficient Data Labelling

In similar cases as previously mentioned, when there is a lack of sufficient labelled data and it is expensive to generate large amounts of data for training a DL model, AL techniques can be used. AL techniques are different from TL techniques as they provide a platform for the model to iteratively learn from new, informative data that can improve its performance. There is a broad category of stream-based and pool-based sampling methods out of which the pool-based strategy is used here.

When the unlabelled data samples arrive periodically in the form of streams, it is impractical to accumulate all the unlabelled samples into a pool to be labelled at a future time (Cheng et al., 2013). In this scenario, the stream-based approach is used wherein the learner makes a decision if the incoming unlabelled sample must be labelled or not. In a pool-based scenario, the unlabelled pool is fixed and is queried exhaustively in every iteration to select the most informative samples.

In pool-based AL, the available data is divided into a labelled pool and an unlabelled pool. The labelled pool contains annotations and is used to train the model which is then evaluated on all the data in the unlabelled pool to predict the most informative samples which would help the model improve, if its labels are provided for training the model. The informative samples are the ones that the model is most uncertain about.

The informative samples can be queried using various uncertainty sampling techniques such as least confidence, margin of confidence and entropy (Settles, 2009). The samples with the highest rank corresponding to the highest entropy are handpicked and their labels are disclosed to the model to learn from. The samples selected through the query are labelled, usually by a human annotator, called the oracle and this data is then added to the training set (labelled pool) on which the model is re-trained and further observed for performance improvements. The selected data sample is then removed from the unlabelled pool.

Different query techniques exist that help in selecting the data points that would help the model the most to improve its performance. The different query types that can be employed in the pool-based AL method include: margin sampling, entropy sampling and least confidence.
In the least confidence method, the single class probability of the predicted class is considered. If the model (with parameters $\theta$) is not confident about the class prediction, such that the class probability of the output class prediction is not closer to 1, then these samples result in high uncertainty for the model. It can be computed as follows:

$$x_{LC}^* = \text{argmax}_x 1 - P_\theta(y|x)$$ (2.2)

where $y$ represents the predicted class i.e. the class label with the highest posterior probability for the most probably class. This method considers only the most probable class and discards further information regarding the other probability of other labels.

The margin sampling method considers the difference between the most likely and next most likely class prediction in terms of posterior probability. If the model is not confident with the prediction, then a small margin between the most likely and the next most likely class predictions are observed. These are the samples that the model is most uncertain about. It can be computed as follows:

$$x_{M}^* = \text{argmin}_x P_\theta(y_1|x) - P_\theta(y_2|x)$$ (2.3)

where $y_1$ and $y_2$ represent the classes with the highest posterior probabilities. Instances in which there is a small margin between the top two most likely classes, are considered the most ambiguous samples. Although this method is an improvement over the least confidence method, it still does not consider the entire output class probability distribution in determining the informative samples which is a drawback for datasets that contains a high number of labels.

Entropy sampling is the technique in which the per-pixel entropy of every pixel in the image of the unlabelled pool is calculated and can be easily applied to multi-label problems. The pixel entropy values are added across the individual image and the images with the highest entropy values are regarded as the most informative samples and are provided to the model, along with their annotations, for learning. This is the case because
the model is most uncertain about these images and they would help the model improve if they were made available. The per-pixel entropy is calculated over the output class probability distribution from softmax as follows:

\[ x^*_E = - \sum_{i=1}^{n} P_\theta(y_i|x) \log P_\theta(y_i|x) \]  \hspace{1cm} (2.4)

The image entropy is simply a sum of all the pixel entropies in the image. In this thesis work, the entropy measure is employed because it considers the entire class probability distribution and does not favour the samples for which the labels is highly likely (SETTLES, 2009). In this method, the top-k images with highest entropy are selected in every iteration and provided for model training. Also, we perform batch queries to select a small batch of samples per query, in order to save time and computational resources that would otherwise be required in the case of selecting a single sample per query. The pixel entropy calculation is inspired by the approach in (FIGUEROA et al., 2012; SIDDIQUI et al., 2020) for pixel scoring and label acquisition.

**Algorithm 2** Active Learning for Efficient Data Labelling (AL)

Given:

- \( x_{L_i}, y_{L_i} \) where \( x_{L_i} \in X_L \) and \( y_{L_i} \in Y_L \) in \( L = \{(x_{L_1}, y_{L_1}), \ldots, (x_{L_n}, y_{L_n})\} \)
- \( x_{U_i} \) where \( x_{U_i} \in X_U \) in \( U = \{(x_{U_1}, \ldots, x_{U_n})\} \)

where \( 0 < n_L << n_U \)

Model Training and Query:

while \( i < N_{iter} \)

i. Train a model \( f_L(\cdot, \theta) : X_L \rightarrow Y_L \) with parameters \( \theta \)
ii. Predict on all samples in unlabelled pool \( U \) with \( f_L(\cdot, \theta) \)
iii. Calculate per-image entropy (2.4) on all predictions from \( U \)
iv. Retrieve top \( E \) highest entropy samples from iii
v. Label the samples selected in iv though the oracle:

\( U \leftarrow U - E \)
\( L \leftarrow L + E \)
\( i \leftarrow i + 1 \)

end while
where $L$ represents the labelled pool with data samples with annotations provided. $U$ represents the unlabelled pool which only contains unlabelled samples such that $0 < n_L << n_U$, wherein the number of labelled samples $n_L$ is much smaller in comparison to a large pool of unlabelled samples $n_U$. $E$ represents the number of samples to be selected in every iteration, which is configurable. The number of iterations can be pre-determined or the process can be iteratively performed until a satisfactory model performance is achieved.

In order to overcome the bottleneck of huge amounts of required training data which are expensive to obtain, this research utilises uncertainty-based sampling for the active selection of informative samples to be labelled (Holub et al., 2008; Joshi et al., 2009). In this these papers, 5 uncertain samples are chosen per round and provided for training. These works compare the different pool-based sampling techniques and it was observed that entropy-based AL outperformed random sampling which was used for passive learning.

### 2.5.1 Advantages of employing Active Learning

Although it becomes computationally expensive to query very large datasets to choose the most uncertain samples, advantages exist while the dataset is within the reasonable limits of the available computational resources.

- Since the model learns the true labels of the uncertain samples, it makes the model learn the decision boundaries better and makes the model more confident in its predictions.

- AL is beneficial as it helps the model to learn incrementally and interactively through human-in-the-loop with small inputs from human users of the system or domain experts.

- It aims to achieve high performance gains with a minimal dataset by proactively selecting the most important and informative samples for the model to learn from. This helps the model to reach high levels of performance without the need to label all the images in the dataset. Hence, eliminating the need for tedious labelling efforts.
2.6 Evaluation Metrics

All the deep-segmentation models trained as a part of this thesis work are evaluated based on their Jaccard index or mIoU value. IoU is used to calculate the percentage overlap in the number of pixels between the actual and predicted segmentation masks, and it can be defined as:

$$\text{IoU}(target, prediction) = \frac{(target \cap prediction)}{(target \cup prediction)}$$ \hspace{1cm} (2.5)

In this case of multi-class SS, the IoU values are calculated individually for each class and averaged over all the classes to calculate the mean IoU per image.

Here, IoU is preferred to pixel accuracy as it is observed that IoU performs better in cases where class imbalance exists. This is attributed to the increase in pixel accuracy even though the minority class instances are incorrectly classified. As we deal with a variety of classes such as buildings and background spanning many more pixels when compared to the area covered by small shrubs, in the green cover class, we can observe an imbalance in class distribution.
3.1 Research Questions

This thesis aims at addressing the following research questions with respect to semantic segmentation in aerial imagery:

- **RQ1**: How can we address domain shift with transfer learning for the purpose of semantic segmentation and label augmentation?

- **RQ2**: How does the performance of transfer learning compare with active learning to achieve efficient multi-class semantic segmentation?

- **RQ3**: How to teach transfer learning-based deep segmentation models with active learning for efficient performance with a limited dataset?

RQ2 and RQ3 are novel approaches as previous works for deep-segmentation in aerial imagery do not consider a comparison and combination of active learning and transfer learning techniques for domain shift and model improvement.
3.2 Methods

Currently, the ObViewSly dataset provides 'building' annotations and this dataset has been utilized for training deep-segmentation models that are capable of detecting buildings. We would like to augment this feature-set with additional labels such as roads and green cover while comparing the AL and TL approaches for the same. The core idea is to use AL to automatically select the fine-tuning data for the TL model. In further discussions, U-Net and Shuffle-Unet are interchangeably referred to as Unet and S_Unet respectively.

3.2.1 Dataset Description and Domain Shift

Figures 3.1 and 3.2 depict a 256x256 patch of both the ObViewSly and the LandCover.ai datasets. As described, in both the datasets, the inputs $x_i$ belong to the 3 channel RGB color space where each channel has values in the range of 0-255. The per-pixel annotation $y_i$ belongs to 1 of the 4 classes ie. building, road, green cover or background.

The ObViewSly dataset has a high spatial resolution of 10cm GSD whereas the open source LandCover.ai dataset and Massachusetts road and building datasets that contain several annotations of interest; available for the SS task, possesses a GSD in the range of 20cm - 50cm and 100cm respectively, causing a domain shift in the source and target domain data as seen in figures 3.1 and 3.2. This leads to a difference in the granularity in the semantic objects captured in the images of the source and target domain. The source domain ie. LandCover.ai dataset can be described as follows:

$$D_S = \{(x_{S_1}, y_{S_1}), \ldots, (x_{S_n}, y_{S_n})\}$$ and $25cm \leq GSD \leq 50cm$

where

$$x_{S_i} \in X_S \subseteq \mathbb{R}^{256x256x3}$$

$$y_{S_i} \in Y_S \subseteq \mathbb{R}^{256x256x4}$$

The target domain ie. the ObViewSly dataset can be described as follows:

$$D_T = \{(x_{T_1}, y_{T_1}), \ldots, (x_{T_n}, y_{T_n})\}$$ and GSD = 10cm

where
There are 4 labels present in the LandCover.ai and ObViewSly datasets, namely: *roads*: depict all the paved and unpaved roads. *buildings*: represent all residential, commercial and industrial buildings. *green cover*: depicts woodlands, i.e. trees and large shrubs/hedges. *background*: includes all other ground objects such as water, agricultural land, grass and others.

The 4 different classes are represented in the segmentation map as follows:

- class 0: Background = black
- class 1: Road = grey
- class 2: Building = brown
- class 3: Green cover = green
3.2.2 Active Learning vs. Transfer Learning for Semantic Segmentation in Aerial Imagery

Both AL and TL provide excellent alternatives in cases where we would like to train data-hungry DL models with a small dataset of available annotations. Through experimentation, we would like to determine if either of these techniques individually or in combination would be a better alternative for our particular use case.

Transfer Learning

In order to address RQ1, with the help of models that can identify these features without the tedious manual effort of exhaustively labelling the ObViewSly dataset i.e. the target domain, transfer learning methods are employed. In order to implement a TL model that is capable of efficient feature extraction in the target domain, the main drawback observed is the lack of labelled datasets that have the same characteristics as that of the
ObViewSly dataset, to perform transfer learning for direct feature extraction from the target domain. We ensure the source and target tasks $T_S = T_T$ are SS in aerial imagery and it closed set as $Y_S = Y_T$ i.e. the label-set is the same in both the source and target task.

In order to facilitate the domain adaptation between these datasets and to train a deep-segmentation model that is capable of identifying different features in the ObViewSly dataset, we employ the transfer learning technique wherein only the final convolution-reLU block and output layer are fine-tuned. In the trained, U-Net-based models, $f_S(\cdot, \theta)$ all the layers except the final convolution-reLU block and output layer are frozen, such that its representations learnt from the source domain remain unchanged, and the unfrozen layers are re-trained/fine-tuned with a limited set of manually labelled samples from the ObViewSly dataset (target domain) to produce a model $f_T(\cdot, \theta_T)$. This model is now competent to identify previously non-existent segments, such as roads and trees, in the ObViewSly dataset and can hence be augmented to the existing feature-set.

Initially, both the Massachusetts and Landcover.ai datasets were considered for TL. But, due to the diversity of annotations, larger size of the dataset and better quality of TL results from the LandCover.ai dataset, it is considered further for comparison with AL. As stated by (Mnih, 2013), it is more valuable to use small data samples that pertain to the task when compared to training a model with large volumes of unrelated data samples. To this end, we are building and training custom deep-segmentation models on relevant datasets and using them to perform transfer learning as opposed to utilising models pre-trained on generic datasets.

**Active Learning**

AL is chosen as one of the techniques as we can observe the DL model's learning behaviour as new samples are iteratively provided. Initially, a random set of 1000 ObViewSly samples are selected such that they are representative of the population (urban/suburban areas) and they are manually labelled by us to generate their respective annotations. A model $f(\cdot, \theta)$ is learnt on the labelled dataset. The AL process also results in SS and further enhancement in the feature-set of the ObViewSly data.

In the AL experiments, there are several query strategies available, and
we have chosen the ‘Entropy’-based strategy to retrieve the most informative samples from the unlabelled pool (as discussed previously) to promote more confident predictions at the pixel level by entropy minimization (Toldeo et al., 2020). When there is a tie among several samples with same entropy values, tie-breaking is performed randomly. The chosen samples are then manually labelled and appended to the existing labelled dataset and removed from the unlabelled pool for the next iteration of model training and evaluation.

Querying entails the selection of images wherein different works select different number of samples per query in the batch mode AL such as (Joshi et al., 2009) which selects 5 images per query and other works perform stream-based AL. In this work, 10 samples are chosen per query such that 25 queries are performed per model and this adheres to the time and computational resources available.

In order to address RQ2, we start with 1000 random 256x256 patches of ObViewSly dataset and reach 1250 patches through the AL process, in order to compare its performance with the corresponding TL model that is fine-tuned on 1250 random ObViewSly image patches.

In the AL pipeline, the models are trained from scratch in every iteration in order to observe the importance of the chosen samples which is manifested through the model performance.

**Active Learning guided Transfer Learning (TL+AL)**

In order to address RQ3, we combine AL with TL. These two approaches ie. AL and TL are consolidated in order to iteratively query informative samples from the target domain ie. ObViewSly dataset during the domain adaptation process, to teach the TL model. Hence, the TL model is fine-tuned with the samples chosen by AL. This is carried out in order to identify the data samples that are highly informative from the target domain that would most likely improve model performance and to evaluate the performance of this combined approach with the individual techniques of active and transfer learning respectively.

This research is performed in order to test if high performance gain is obtained, with a limited dataset, in the combined approach as the individ-
ual approaches are both geared towards the training of high-performing deep-segmentation models from a limited dataset.

**Algorithm 3** Active Learning guided Transfer Learning TL+AL

**Given:**
\[ x_{S_i}, y_{S_i} \text{, where } x_{S} \in X_S \text{ and } y_{S} \in Y_S \text{ in } D_S = \{(x_{S_i}, y_{S_i})\}_{i=1}^n \]
\[ x_{TL_i}, y_{TL_i} \text{ where } x_{TL} \in X_{TL} \text{ and } y_{TL} \in Y_{TL} \text{ in } L = \{(x_{TL_i}, y_{TL_i})\}_{i=1}^n \]
\[ x_{TU_i} \text{ where } x_{TU} \in X_{TU} \text{ in } U = \{(x_{TU_i})\}_{i=1}^n \]
where \( 0 < n_{TL} << n_{TU} \) and \( 0 < n_{TL} << n_S \)

**Model Pre-training:**
Learn a model \( f_S(\cdot, \theta) : X_S \rightarrow Y_S \) with parameters \( \theta \)

**Initial Model Fine-tuning:**
Fine-tune the unfrozen parameters in \( f_S(\cdot, \theta) : X_S \rightarrow Y_S \) to learn a model
\( f_T(\cdot, \theta_T) : X_{TL} \rightarrow Y_{TL} \) with fine-tuned parameters \( \theta_T \)

**AL Query+Fine-tuning:**
while \( i < N_{iter} \)
  i. Predict on all samples in unlabelled pool U with \( f_T(\cdot, \theta_T) \)
  ii. Calculate per-image entropy (2.4) on all predictions from U
  iii. Retrieve top 10 highest entropy samples from ii
  iv. Label the samples selected in iii through oracle:
      \[ U \leftarrow U - E \]
      \[ L \leftarrow L + E \]
  v. Fine-tune the unfrozen parameters in \( f_T(\cdot, \theta_T) : X_{TL} \rightarrow Y_{TL} \)
      \[ i \leftarrow i + 1 \]
end while

By combining the two approaches, we can address both the tasks of domain shift as well as the iterative fine-tuning of the model with a small dataset. In this approach, the utility of the data samples in the target domain can be determined and utilised during the domain adaptation process with transfer learning. As an example, an approach was proposed (after the inception of this thesis work), that combines both AL and TL with the aim of training a learner with a significantly small training dataset for brain tumour classification where it was shown that only 30% to 60% of the dataset was required to be annotated in the target domain for fine-tuning
the model (Hao et al., 2021). This is one of the early works in this field and has only been applied in the medical domain so far.

This concept is demonstrated in 3.3 for SS in aerial imagery wherein the Unet-based model is divided into 9 blocks and only the final block i.e. block 9 and the output layer are considered in the fine-tuning process. The unlabelled pool is queried, manually annotated and appended to the labelled pool to be used for model fine-tuning. This process can be formally described with the pseudocode 3. Here, the source domain \( D_S = \{(x_{S_i}, y_{S_i})\} \) consists of pairs \((x_{S_i}, y_{S_i})\) where \(x_{S_i} \in X_S\) is a data instance in the LandCover.ai dataset and \(y_{S_i} \in Y_S\) is the corresponding label, and \(f_S(\cdot, \theta)\) is the objective function that maps \(X_S \rightarrow Y_S\) such that \(D_S = \{(x_{S_1}, y_{S_1}), ..., (x_{S_n}, y_{S_n})\}\). Similarly, ObViewSly is the target domain \(D_T\) that is made up of a labelled pool \(L\) that contains \(x_{TL_i} \in X_{TL}\) and its corresponding labels \(y_{TL_i} \in Y_{TL}\) and an unlabelled pool \(U\) that contains samples \(x_{TU_i} \in X_{TU}\). The number of iterations for fine-tuning the TL model with AL can be pre-determined or the process can be iteratively performed until a satisfactory model performance is achieved. In \(D_T\), the number of labelled samples is much smaller

---

**Figure 3.3: Active Learning guided Transfer Learning**

---

---
3.2. METHODS

than the number of unlabelled samples ie. $0 < n_{TL} << n_{TU}$ and the number of available annotated data samples in the source domain is much larger than that in the target domain ie. $0 < n_{TL} << n_{S}$.

3.2.3 Proposed Model Architecture

The above methods to address RQ1-3 are applied on U-Net to provide a baseline performance. Then, the Shuffle-Unet model (based on the original U-Net structure) is trained using the above mentioned techniques and the performance of the two models in the transfer learning and active learning settings are evaluated. Additionally, due to deconvolution overlap and the resulting checkerboard artifacts, in the original U-Net model, observed during the preliminary experimentation process, bilinear upsampling is used followed by 2x2 convolutions in the expansion path similar to the study in (AMIRI et al., 2020).

Shuffle-Unet

The Shuffle-Unet, as shown in figure 3.4 is the proposed, experimental DL model based on the original U-Net model with modifications such that the maxpooling and upsampling layers are replaced by the periodic unshuffle and periodic shuffle layers respectively as discussed in 2.3.2. Hence, the U-Net pixel shuffle model helps to retain information present in the image during downsampling, in order to learn better representations and to generate checkerboard artifact-free segmentation maps in the upsampling path. At every downsampling stage with periodic unshuffle, with a downsampling factor of 2, the number of channels increase to $c \times 2^2$. In order to reduce the channels to reflect the dimensions in that of the original U-Net, 1x1 convolutions are employed. The same strategy is used in the upsampling path where channels are reduced to $c \div 2^2$ at every periodic shuffle. Here, 1x1 convolutions increase the number of channels in the segmentation maps. Finally, the output segmentation map generated by the model is passed through the softmax function to obtain per-pixel class probability distribution. The Shuffle-Unet model architecture is shown in the figure 3.4. The Shuffle-Unet model and the original U-net (with modifications) are employed in this thesis work to address RQ1-3 for SS and label augmentation in aerial imagery.
Figure 3.4: The Shuffle-Unet Architecture
Prototype Implementation

The implementation is performed in different stages and widely encompasses the following tasks:

- Dataset Preparation
- Model Implementation

4.1 Dataset Preparation

The focus of this stage is to perform data labelling, and dataset preparation. The data labelling process encompasses the generation of the pixel annotations and the survey of different annotation tools for doing so. Approximately 15 minutes were required to label every 256x256 patch.

4.1.1 Data Labelling and Augmentation for the ObViewSly Dataset

1. An initial dataset consisting of approximately 2000 images of 1024x1024 pixels was chosen from the ObViewSly dataset.

2. Initially, a selection of 2500 patches of 256x256 pixels was made to be used for training, validation and testing of the model, and it was ensured to select a diverse set of samples from both densely and sparsely populated regions.

3. From these selected 256x256 samples, pixel-level annotations were generated for both roads and green cover. These annotations were combined with the existing annotations for buildings. Hence the prepared dataset consists of pixel classes: roads(1), buildings(2), green
cover(3), background(0). Different annotations for each image were generated separately and combined together before the model training process. This was done to provide flexibility for adding/removing annotations in the future.

4. Augmentations such as vertical flip, horizontal flip, increase and decrease in brightness, cropping and zooming were performed to increase the size of the dataset by 7 times. These 256x256 patches of RGB data is normalized to a range of 0-1 by dividing with a value of 255 in order to make the computations for model learning faster and easier.

Different tools were surveyed during the process of manual label preparation. These include tools such as LabelIMG, labelme, VoTT, imglab and several others. ImageJ and 3D Slicer were chosen as they allowed the direct export of the generated masks in the form of images and allowed to perform the annotation locally without the need for uploading the images online. They also provided sufficient ease of use for producing the annotations.

4.1.2 Data Preparation for the Massachusetts Road and Building Datasets

1. Massachusetts road dataset: The original dataset of road annotations consists of 1500x1500 px images which were cropped into a size of 256x256 patches for training the CNN model. Massachusetts building dataset: The dataset consists of pixel annotations for buildings with similar dimensions as the road dataset.

2. The two datasets were combined to form a dataset with both road and building annotations and this was used in the model training process in patches after normalization.

4.1.3 Data Preparation Process for the LandCover.ai Dataset

The LandCover.ai dataset consists of several GeoTiff tiles that span suburban and rural areas. The regions of interest in accordance to the ObViewSly dataset corresponded to 5 tiles namely N-33-130-A-d-4-4.tif, N-34-61-B-a-1-1.tif, N-34-97-C-b-1-1.tif, N-34-140-A-d-3-4.tif and N-34-140-A-d-4-2.tif.
1. Initially 5 classes were provided with the corresponding labels: background(0), building (1), woodland (2), water(3), road(4). These labels were transformed to labels: building (2), woodland (3), road(1) and background(0) in order to conserve the class numbers over all the experiments. All the available labels except ’Water’ was used as they were found to be more suitable for the socio-economic analysis of the region and hence were considered for the experiments.

2. The data was then augmented with random horizontal/vertical flip cropped into patches of 256x256 and was normalized to a range of 0-1 for model training.

### 4.2 Model Implementation

As the focus of the thesis work is to compare the active learning and transfer learning techniques, we have used Encoder-Decoder based models such as Unet and Shuffle-Unet. In total, around 75 experiments of the models training and evaluation, apart from other preliminary experiments, were performed. Each model required approximately 6 to 8 hours to complete.

#### 4.2.1 Hyperparameter Tuning

In order to study the effect of different hyper parameters such as learning rate, batch size and filter size on the domain shift problem (MNIH, 2013), several experiments were performed with a combination of grid search with random search. Several hyper-parameter combinations of learning rates of 0.0001, 0.0003, 0.0005 and 0.001, batch size of 8, 32 and 64, and two different filter sizes of 3x3 and 7x7 were performed for the two models on the ObViewSly dataset. Learning rate 0.0001, batch size 8 and filter size of 3x3 outperformed the other combinations that were tested and were used as the hyperparameters in all the experiments.

#### 4.2.2 Technical Details

Pytorch was used to build, train and evaluate the models. The experiments were conducted on several GPUs as per their availability. These included Tesla V100-SXM2, T4, NVIDIA GeForce RTX 2080, Quadro RTX
4000 and GeForce GTX 1050 Ti with sizes ranging from 4GB to 32GB.

### 4.3 Model Training and Evaluation

A 70-30-10 split of the data for training, validation and testing, was used throughout this work. All these 256x256 patches were randomly rotated by 90°, 180°, 270° or chosen not to be rotated randomly. This is done in to reduce the orientation bias that could be induced in the deep-segmentation model (NIH, 2013). For uniformity, the models were trained for 150 epochs with an early stopping regularization criterion wherein the training process was terminated if the patience duration of 20 epochs were exceeded without any reduction in validation loss. Further details include the use of Adam optimizer with a learning rate 1e-4 with the categorical cross-entropy loss function and a batch size of 8. Every experiment was initialised with the same random seed value to facilitate reproducibility in model training. The model outputs are passed through the softmax activation function to obtain the class probability distribution per pixel. The trained models were evaluated with the help of their mIoU values on the test set.

### 4.4 Experiments

All the experiments for TL, AL and a combination of TL and AL as described in 3.2.2 are performed for both the U-Net and Shuffle-Unet architectures.

**Transfer Learning**

The Transfer learning experiments were conducted in multiple stages. Initially, the Massachusetts and LandCover.ai datasets were used to train Unet and Shuffle-Unet. The models were then fine-tuned using the annotations from 1250, randomly chosen, patches from the ObViewSly dataset. The learnable parameters of all but the last convolutional block and output layer of the pre-trained model were frozen, and the parameters of the unfrozen layers were fine-tuned.
Active Learning

For the active learning technique, the training process began with 1000 randomly chosen 256x256 patches from the ObViewSly dataset. Then, the model was used to predict on the unlabelled pool of data in order to determine the most uncertain samples which would provide the most information to the model. The entropy-based query strategy was used wherein the pixel-wise entropy was calculated and summed across the entire image. 10 images with the highest entropy was chosen for manual labelling. The 10 samples are labelled, augmented, normalized (as described in 4.1.1) and then added to the labelled pool, removed from the unlabelled pool and the model was then re-trained on the augmented labelled pool. The model training is done from scratch in every iteration so as to observe the effect of the selected images on the model’s performance. The query process is repeated for 25 iterations and the model performance on individual class IoU and mIoU are evaluated.

Active Learning guided Transfer Learning

As described in 3.2.2, the TL model is provided informative samples based on entropy values and 10 queries are performed for fine-tuning the TL models with AL, which are iteratively evaluated.
All the experiments conducted in order to address the research questions are evaluated here. The performance of different models with the TL, AL and AL+TL techniques are compared to understand if a particular approach performs better on the SS task on aerial imagery with the ObViewSly dataset. The evaluation results of the experiments are presented and discussed in this chapter in terms of several measures such as IoU, mIoU, Entropy and other details are provided.

5.1 Evaluation

The results are evaluated considering the mIoU between 0.5 and 1.0. The lower threshold of 0.5 is used to indicate that IoU < 0.5 does not have sufficient overlap between the ground truth and predictions and are hence not considered as a part of this analysis.

The predictions of the pre-trained Unet and S_Unet TL models on the ObViewSly dataset (directly without fine-tuning) yielded IoU < 0.5 in both models. Hence the decision was made to fine-tune these TL models by freezing the learned representation in all the layers except the last conv-relu block and the output layer as described.

Sample RGB images from the test set along with their corresponding ground truth and segmentation maps generated by the Unet and S_Unet for all the three techniques of TL, AL and TL+AL are provided below in figures 5.1 and 5.2.
Figure 5.1: Unet - Comparison of TL, AL and TL+AL on the test set
5.1. EVALUATION

Figure 5.2: S_Unet - Comparison of TL, AL and TL+AL on the test set
The performance evaluation of the Unet and S_Unet models on the Ob-ViewSly dataset for the AL, TL and TL+AL techniques in comparison with their performance while training on 1250 random samples, are represented in terms of mIoU in figures 5.3 and 5.4. The AL approach starts with 1000 random samples as shown on the x-axis and the resulting mIoU obtained by training the model by iteratively adding the 10 highest-entropy (most informative) samples is shown on the y-axis. In each AL run, the model is trained from scratch in order to observe the effect in performance (impact) caused by the high entropy samples. mIOU is calculated per image and the graph represents the average mIOU of the test set on different runs of the models. Further, the distribution of mIoU and per-class IoU over the different runs of AL and TL+AL are provided in the subsequent plots 5.5 and 5.6.

The S_Unet model undergoes a performance degradation when the last layers of the pre-trained model are fine-tuned on 1250 random samples in TL (S_Unet-TL). It also exhibits slightly lower performance when trained from scratch on the same collection of 1250 random samples (S_Unet-Random). Some performance improvement is observed in several AL iterations of the Unet (Unet-AL). At the end of 25 AL iterations that chose high entropy samples for training, the Unet trained on 1250 (high entropy) samples had mIoU score comparable to that which was trained on 1250 random samples from scratch without AL (Unet-Random). Therefore, in Unet, random selection had similar performance as that of AL with entropy query strategy. Unet also indicates a better TL performance when the pre-trained model is fine-tuned on 1250 random samples.

Further, the high entropy samples are the regions of interest as these are the samples for which the model is unsure of the per-pixel classification and are demonstrated in this section along with their corresponding entropy heatmaps. The heatmaps clearly depict the regions of the image that the model is unsure about as seen in 5.7. Here, some examples of low entropy images for which the model is certain are also shown. Further, the entropy values over all the samples in the entire unlabelled pool for the last AL run can be observed for both the Unet and S_Unet models in figures 5.8 and 5.9. These observations also provide interesting insights into the behavior of different models while employing the same training techniques.
5.1. EVALUATION

Figure 5.3: AL vs TL mIoU performance

Figure 5.4: AL guided TL mIoU performance
Figure 5.5: Distribution of mIoU in the test set across all the AL and TL+AL runs

Figure 5.6: Distribution of individual class IoU in the test set across all the AL and TL+AL runs
5.1. EVALUATION

Figure 5.7: Examples of high and low entropy samples observed in the AL rounds
Figure 5.8: Unet Entropy distribution across all the samples in the unlabelled pool

Figure 5.9: S_Unet Entropy distribution across all the samples in the unlabelled pool
5.2 Discussion

We proved that it is possible to perform domain adaptation in RS image data that vary in GSD even where $GSD_S > 2GSD_T$ i.e. the GSD in the source domain is more than twice the GSD in the target domain, with the help of TL. Fine-tuning of pre-trained models through TL for the task of SS in aerial imagery was performed similar to the TL demonstrated by several other studies (Perello und Kuffer, 2020; Ulmas und Liiv, 2020). Extensive research was performed on the availability of different RS datasets, and Massachusetts and LandCover.ai datasets were selected for the purpose of pre-training. Also, we have successfully labelled around 2500 image samples manually in order to facilitate the training of the DL models.

AL, an alternative to TL, was seen to provide comparable performance with a smaller set of annotated samples in the case of S_Unet. Also, it can be observed that there is an overall improvement in performance in the AL iterations for Unet. With respect to overhead in terms of time and computational resources, AL required much higher training times due to the several iterations but it was a trade-off as we could observe a visible increase in performance on some runs. For the TL process, a large amount of time was necessary to research and evaluate several datasets, and to manually generate ground truth annotations for model fine-tuning.

The main contribution of this work which is our proposed method of AL guided TL proved to work well for Unet wherein the mIoU score was seen to improve when compared to the individual TL or AL runs. In the case of S_Unet, the rugged line in figure 5.4 depicts the mIoU performance initially, but the results towards the end of the 10th iteration of TL+AL look promising. Perhaps, this performance would improve on continued experimentation. The performance of all the proposed techniques, with the intention of addressing our research questions, were evaluated and the following observations were made. Representations learned by periodic Shuffle/Unshuffle operations were not transferable across domains. Hence, S_Unet exhibits low performance in TL across domains and incorrectly classifies some border pixels. Also, since only the last layers were fine-tuned in the TL process, this could have inhibited the model’s performance improvement because S_Unet performed much better than Unet when trained from scratch in the AL pipeline. Similarly, in the AL guided TL pro-
cess, only the last layers were fine-tuned on the ObViewSly dataset which could be the bottleneck that caused weak performance. All the TL, AL and TL+AL techniques were successfully performed by annotating less than 10% of the total initially selected data pool in the ObViewSly dataset. These techniques help to save several 100s of hours and significant manpower and corresponding computational resources.

It could not be ascertained as to which technique is better than the other as it was observed that it is widely dependent on the characteristics of the DL model used for training. Here, it was observed that Unet was more adaptable to TL than S_Unet. In the case of Unet, better performance was observed in TL with the same number of images as those chosen by AL, a behaviour as seen in studies like Amiri et al. (2020). Whereas in the case of S_Unet, it was observed that it lacks the ability to generalize to different domains because it was observed that S_Unet performed significantly better when trained on the target dataset from scratch. This improvement in performance while training from scratch could be due to shuffle-based up-sampling and prevention of information loss that is caused due to pooling. It could be suspected that the degradation in generalization performance of S_Unet could be due to the absence of pooling layers that are responsible for translation invariance to the semantic features present in the image. This behaviour warrants further investigation.

Sometimes, a dip in performance can be observed after an AL run. This could be due to class imbalance that might be introduced when a high number samples of the same class are consistently selected over many AL iterations. Sometimes, due to the lack in diversity of selected samples, the high entropy samples may belong to the same class or have similar semantic features. This might affect the performance of the AL iterations until diverse samples can be selected and added to the labelled pool for training. High entropy samples that were selected for labelling were proven to be those which are difficult for the model to classify such as those samples with high shadow and occlusion effects, samples with blur, problems with illumination and distortions and samples that have a higher number of classes with a distributed green cover and those samples which were previously unseen such as parking lots and large factories.

Entropy values from Unet predictions in the unlabelled pool were much higher in comparison to that of S_Unet but it was observed that over the
5.2. DISCUSSION

course of TL+AL, Unet was able to pick better samples and an increase in mIoU performance was seen. If the model is unsure of the classification of the pixel such as a pixel with the class distribution of 0.1, 0.1, 0.2, 0.6 from softmax, the pixel entropy can reach very high values. This could be the explanation in the case of the entropy distribution observed over Unet as opposed to S_Unet. The entropy values of the top-10 samples picked by the models in each AL and TL+AL run is shown in figures A.2, A.3, A.4 and A.5. Entropy is calculated on the entire predicted class distribution per-pixel and in the case of Unet it was noted that it was able to correctly classify the pixels but the distribution was such that it led to a higher entropy score.

Although these techniques work well despite the wide range of characteristics of brightness, colors and textures of the semantic features, green cover segmentation performed consistently poorly in terms of mIoU, throughout the different models and techniques, although it observed that the models were consistently able to predict "trees" very well. This could be due to the fact that green cover labels are spread out in the image in the form or shrubs/bushes of varying sizes which occupy much less pixel area when compared to larger trees and are usually composed of several different colours, heights and textures and are often victims of shadows and occlusions. This shadow and occlusion effect is usually caused by trees that cover parts of roads, buildings and shrubs or by buildings whose shadows overflow onto the roads and shrubs, making it difficult for the DL model to accurately segment these pixels.

We were able to select a subset of highly informative samples from a large initial dataset. Hence, we were able to perform efficient labelling by selecting the samples that would help the model learn better and to help it to be more sure of the predictions it makes. As stated, it took us around 15 minutes to manually generate annotations for a 256x256 tile. Hence, the time that would be required just to manually annotate, for example, a dataset of 10000 samples would have been 2500 hours, to generate datasets alone. Here, it is proven that data-hungry models can be trained with smaller datasets if the correct techniques are employed.


6

Conclusion and Future Work

6.1 Conclusion

Through this work, we have addressed several research questions such as domain adaption through TL models and its comparison with AL and with our method of AL guided TL. In several iterations, AL improved model performance both when trained from scratch and when it was used to guide TL models. We have successfully implemented the novel and previously unexplored AL guided TL technique for SS in aerial imagery which proves that we can automatically analyse vast expanses of land areas without the need to manually annotate huge datasets. Our methods help to combine the best of both worlds because the models can learn good representations from existing, open-source datasets, and selecting the most informative samples to learn in the target dataset. Thus, it circumvents the need to exhaustively label huge datasets and leads to huge savings in terms of time, money and manpower.

Based on the operations in the model, pixel shuffle operation could not be generalized. Whereas U-Net with pooling layers was able to better transfer the learned representation from the source to the target domains. By visualizing the entropy heatmaps in high entropy images, it was observed that models are affected adversely by regions covered in shadow that make it difficult for the model to predict the underlying objects. Some samples were difficult to learn due to the unusual semantic features that occurred rarely in the dataset such as industrial buildings as opposed to frequently occurring domestic buildings which had visibly different features. We also observed that certain models work better in the AL setting and some in the TL setting based on the operations used in the model. Therefore, the method proposed in this work can be used to build an informative dataset
and a corresponding high performing model regardless of the presence of ground truth annotations.

6.2 Future Work

This thesis work can be applied in several areas which require the understanding of the semantic features present in RS images such as socio-economic analysis. DL methods generally require more data but they are preferred to traditional CV techniques due to their superior performance and feature extraction capabilities. Hence, this work can be applied in all cases wherein DL models can be employed to reduce the annotation effort and to train models that are capable of achieving good performance with small datasets.

Further, it was observed that shadows and occlusions affected the quality of predictions generated by the models. Hence, further research could be done in this area. Also, there was a lack of diversity in high entropy samples in several iterations. In these cases, further research can be undertaken to increase the diversity in the selected samples (Thu et al., 2011). As all the AL runs were sequentially performed once, several repeated runs could be further performed or techniques such as Monte Carlo Simulation can be employed to estimate the highest entropy samples. Also, AL can be carried out for further iterations to observe model performance with increasing number of iteration. Additionally, we could also classify buildings based on their industrial or residential purpose, to gain a deeper understanding of the region and its economy. Depth data can be used in addition to the RGB channel data for the SS process. This would help with semantic features that have varying heights such as that of green cover that encompasses both tall trees and shorter shrubs.

This work has vast applicability in today’s world on account of a never-ending supply of RS image data, accelerated computing capabilities, and the requirement for performing fast-paced, automated analysis in socio-economic niche.
A.1 Detailed results

Figure A.1: Example segmentation maps produced by the different models on the Landcover.ai dataset
## APPENDIX A. DETAILED RESULTS

### LandCover.ai Dataset

<table>
<thead>
<tr>
<th>Roads</th>
<th>Buildings</th>
<th>Green Cover</th>
<th>Background</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unet</td>
<td>0.72</td>
<td>0.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_Unet</td>
<td>0.89</td>
<td>0.94</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.1: mIoU score of Unet and S-Unet on the source-domain during TL model pre-training.

### Massachusetts Road+Building Dataset

<table>
<thead>
<tr>
<th>No. of Images</th>
<th>Roads</th>
<th>Buildings</th>
<th>Green Cover</th>
<th>Background</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1250</td>
<td>0.613</td>
<td>0.567</td>
<td>0.618</td>
<td>0.661</td>
<td>0.613</td>
</tr>
<tr>
<td>1260</td>
<td>0.614</td>
<td>0.614</td>
<td>0.614</td>
<td>0.614</td>
<td>0.614</td>
</tr>
<tr>
<td>1270</td>
<td>0.615</td>
<td>0.566</td>
<td>0.615</td>
<td>0.566</td>
<td>0.615</td>
</tr>
<tr>
<td>1280</td>
<td>0.616</td>
<td>0.567</td>
<td>0.616</td>
<td>0.567</td>
<td>0.616</td>
</tr>
<tr>
<td>1290</td>
<td>0.617</td>
<td>0.568</td>
<td>0.617</td>
<td>0.568</td>
<td>0.617</td>
</tr>
<tr>
<td>1300</td>
<td>0.618</td>
<td>0.569</td>
<td>0.618</td>
<td>0.569</td>
<td>0.618</td>
</tr>
<tr>
<td>1310</td>
<td>0.619</td>
<td>0.570</td>
<td>0.619</td>
<td>0.570</td>
<td>0.619</td>
</tr>
<tr>
<td>1320</td>
<td>0.620</td>
<td>0.571</td>
<td>0.620</td>
<td>0.571</td>
<td>0.620</td>
</tr>
<tr>
<td>1330</td>
<td>0.621</td>
<td>0.572</td>
<td>0.621</td>
<td>0.572</td>
<td>0.621</td>
</tr>
<tr>
<td>1340</td>
<td>0.622</td>
<td>0.573</td>
<td>0.622</td>
<td>0.573</td>
<td>0.622</td>
</tr>
<tr>
<td>1350</td>
<td>0.623</td>
<td>0.574</td>
<td>0.623</td>
<td>0.574</td>
<td>0.623</td>
</tr>
</tbody>
</table>

Table A.2: AL mIoU Values

### Table A.3: TL + AL mIoU Values

<table>
<thead>
<tr>
<th>No. of Images</th>
<th>Roads</th>
<th>Buildings</th>
<th>Green Cover</th>
<th>Background</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unet</td>
<td>0.503</td>
<td>0.516</td>
<td>0.503</td>
<td>0.516</td>
<td>0.503</td>
</tr>
<tr>
<td>S-Unet</td>
<td>0.504</td>
<td>0.517</td>
<td>0.504</td>
<td>0.517</td>
<td>0.504</td>
</tr>
</tbody>
</table>

**Note:** The tables above summarize the mIoU scores for different datasets and model variants, providing detailed results for various image collections.
AL is performed for 25 iterations and TL+AL for 10 iterations with Unet and S.Unet respectively. In every iteration, 10 images with highest entropy are selected. In figures A.2, A.3, A.4 and A.5 the entropy values of the top 10 images selected across all the AL runs are presented.

![Figure A.2: Entropy - Unet AL](image1)

![Figure A.3: Entropy - Unet TL + AL](image2)
APPENDIX A. DETAILED RESULTS

Figure A.4: Entropy - S_Unet AL

Figure A.5: Entropy - S_Unet TL + AL
A.2 Other preliminary experiments

In addition to the experiments on the LandCover.ai dataset, several preliminary experiments were performed on the Massachusetts Road and Building datasets which are discussed here.

![RGB Image - Ground Truth - Unet - S_Unet](image)

Figure A.6: Example segmentation maps produced by the different models on the Massachusetts Road and Building dataset

1. Initially, due to the lack of ground truth annotations required for performing multi-class supervised SS in the ObViewSly dataset, unsupervised clustering was performed using the K-means algorithm by selecting different number of clusters. The results yielded through this method were unsatisfactory.

2. The TL model fine-tuning process was not suitable for the model trained on the Massachusetts dataset as it resulted in mIoU values less than 0.5 when different parts of the model was frozen/fine-tuned on the ObViewSly dataset. But this process resulted in good performance from the TL model pre-trained on the Landocover.ai dataset. It was
also observed that the LandCover.ai dataset is more suited due to similar characteristics of the semantic features as that of the target domain, in terms of size, texture, illumination. Better performance on the target dataset and the availability of a wider range of segmentation annotations also made the LandCover.ai the popular choice when compared to the Massachusetts dataset.
### Abbreviations and Notations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>UAVs</td>
<td>Unmanned Aerial Vehicles</td>
</tr>
<tr>
<td>RS</td>
<td>Remote Sensing</td>
</tr>
<tr>
<td>SS</td>
<td>Semantic Segmentation</td>
</tr>
<tr>
<td>DL</td>
<td>Deep Learning</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>SSL</td>
<td>Semi-Supervised Learning</td>
</tr>
<tr>
<td>CV</td>
<td>Computer Vision</td>
</tr>
<tr>
<td>AL</td>
<td>Active Learning</td>
</tr>
<tr>
<td>TL</td>
<td>Transfer Learning</td>
</tr>
<tr>
<td>CVDL</td>
<td>Computer Vision-based Deep Learning</td>
</tr>
<tr>
<td>IoU</td>
<td>Intersection over Union</td>
</tr>
<tr>
<td>mIoU</td>
<td>Mean Intersection over Union</td>
</tr>
<tr>
<td>YOY</td>
<td>Year-Over-Year</td>
</tr>
<tr>
<td>GSD</td>
<td>Ground Sampling Distance</td>
</tr>
<tr>
<td>OSM</td>
<td>Open Street Maps</td>
</tr>
<tr>
<td>CRS</td>
<td>Coordinate Reference Systems</td>
</tr>
<tr>
<td>UTM</td>
<td>Universal Transverse Mercator</td>
</tr>
<tr>
<td>DSM</td>
<td>Digital Surface Modelling</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural Networks</td>
</tr>
<tr>
<td>CRF</td>
<td>Conditional Random Fields</td>
</tr>
<tr>
<td>DIMD</td>
<td>Data-driven Index of Multiple Deprivation</td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under the ROC Curve</td>
</tr>
<tr>
<td>LR</td>
<td>Low Resolution</td>
</tr>
<tr>
<td>HR</td>
<td>High Resolution</td>
</tr>
<tr>
<td>FCN</td>
<td>Fully Convolutional Network</td>
</tr>
<tr>
<td>TL + AL</td>
<td>Active Learning guided Transfer Learning</td>
</tr>
<tr>
<td>S_Unet</td>
<td>Shuffle-Unet</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>2.1</td>
<td>Massachusetts Road and Building datasets</td>
</tr>
<tr>
<td>2.2</td>
<td>(a) An RGB image from the Landcover.ai dataset (b) Pixel annotations provided in the Landcover.ai dataset</td>
</tr>
<tr>
<td>2.3</td>
<td>U-Net Architecture (RONNEBERGER et al., 2015)</td>
</tr>
<tr>
<td>2.4</td>
<td>Periodic Shuffle operation (AITKEN et al., 2017)</td>
</tr>
<tr>
<td>3.1</td>
<td>(a) Sample 256x256 RGB inputs (b) Corresponding manually annotated ground truth from the ObViewSly dataset</td>
</tr>
<tr>
<td>3.2</td>
<td>(a) Sample 256x256 RGB inputs (b) Corresponding ground truth provided in the LandCover.ai dataset</td>
</tr>
<tr>
<td>3.3</td>
<td>Active Learning guided Transfer Learning</td>
</tr>
<tr>
<td>3.4</td>
<td>The Shuffle-Unet Architecture</td>
</tr>
<tr>
<td>5.1</td>
<td>Unet - Comparison of TL, AL and TL+AL on the test set</td>
</tr>
<tr>
<td>5.2</td>
<td>S_Unet - Comparison of TL, AL and TL+AL on the test set</td>
</tr>
<tr>
<td>5.3</td>
<td>AL vs TL mIoU performance</td>
</tr>
<tr>
<td>5.4</td>
<td>AL guided TL mIoU performance</td>
</tr>
<tr>
<td>5.5</td>
<td>Distribution of mIOU in the test set across all the AL and TL+AL runs</td>
</tr>
<tr>
<td>5.6</td>
<td>Distribution of individual class IoU in the test set across all the AL and TL+AL runs</td>
</tr>
<tr>
<td>5.7</td>
<td>Examples of high and low entropy samples observed in the AL rounds</td>
</tr>
</tbody>
</table>
5.8  Unet Entropy distribution across all the samples in the unlabelled pool .............................................. 54
5.9  S_Unet Entropy distribution across all the samples in the unlabelled pool .............................................. 54

A.1  Example segmentation maps produced by the different models on the Landcover.ai dataset .......................... 61
A.2  Entropy - Unet AL ................................................................. 63
A.3  Entropy - Unet TL + AL ...................................................... 63
A.4  Entropy - S_Unet AL ............................................................. 64
A.5  Entropy - S_Unet TL + AL .................................................... 64
A.6  Example segmentation maps produced by the different models on the Massachusetts Road and Building dataset .......................... 65
List of Tables

2.1 Survey of aerial datasets and deep learning applications . . . 8

A.1 mIoU score of Unet and S_Unet on the source-domain during TL model pre-training . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 62
A.2 AL mIoU Values . . . . . . . . . . . . . . . . . . . . . . . . . . . 62
A.3 TL + AL mIoU Values . . . . . . . . . . . . . . . . . . . . . . . . 62
## List of Algorithms

1. Domain Adaptation with Transfer Learning (TL) .......................... 23
2. Active Learning for Efficient Data Labelling (AL) ......................... 28
3. Active Learning guided Transfer Learning TL+AL .......................... 37

[AITKEN et al.] Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize.


[Ronneberger et al.] U-Net: Convolutional Networks for Biomedical Image Segmentation.


Declaration of Academic Integrity

I hereby declare that I have written the present work myself and did not use any sources or tools other than the ones indicated.

Datum: .................................................................

(Signature)