Unsupervised image segmentation in satellite imagery using deep learning

Suraj Regmi

Follow this and additional works at: https://louis.uah.edu/uah-theses

Recommended Citation
https://louis.uah.edu/uah-theses/458

This Thesis is brought to you for free and open access by the UAH Electronic Theses and Dissertations at LOUIS. It has been accepted for inclusion in Theses by an authorized administrator of LOUIS.
UNSUPERVISED IMAGE SEGMENTATION
IN SATELLITE IMAGERY USING DEEP LEARNING

Suraj Regmi

A THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in
The Department of Computer Science
to
The Graduate School
of
The University of Alabama in Huntsville
May 2023

Approved by:
Dr. Huaming Zhang, Research Advisor/Committee Chair
Dr. Manil Maskey, Committee Member
Dr. Chaity Banerjee Mukherjee, Committee Member
Dr. Letha Etzkorn, Department Chair
Dr. Rainer Steinwandt, College Dean
Dr. Jon Hakkila, Graduate Dean
Abstract

UNSUPERVISED IMAGE SEGMENTATION IN SATELLITE IMAGERY USING DEEP LEARNING

Suraj Regmi

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Department of Computer Science
The University of Alabama in Huntsville
May 2023

Image segmentation is typically done through supervised learning. Supervised learning requires labeled data, which is costly and time-consuming to acquire. However, unlabeled data is abundant. This research presents an application of the state-of-the-art unsupervised instance segmentation method FreeSOLO in satellite images and benchmarks the method in iSAID, CrowdAI, and PASTIS datasets. The method achieved 0.9% $AP_{50}$ in the iSAID dataset, 3.1% $AP_{50}$ on the CrowdAI dataset, and 1.1% $AP_{50}$ on the PASTIS dataset. On large objects, it achieved 1.2% $AP_{50}$ in the iSAID dataset and 3.5% $AP_{50}$ in the CrowdAI dataset. The method was also tested in the UAH periphery and MBRSC Dubai dataset where the model was able to segment buildings, water bodies, highways, apartments, and trees. This research also demonstrates the comparative performance of FreeSOLO-based weights relative to other popular supervised learning-based encoder weights on semantic segmentation downstream task.
Acknowledgements

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Dr. Huaming Zhang, for his guidance, support, and encouragement throughout the duration of this project. I am deeply grateful for his invaluable insights, constructive feedback, and patience, which have been instrumental in shaping the direction and quality of my research. I would also like to express my gratitude to the members of my advisory committee, Dr. Manil Maskey and Dr. Chaity Banerjee Mukherjee for their valuable feedback, guidance, and support throughout the development of this thesis. Their insights and expertise have been invaluable in helping me to refine my research questions, focus my analysis, and improve the overall quality of my work.

I am also grateful to the University of Alabama in Huntsville and my GRA team, NASA-IMPACT, for providing me with the resources and support necessary to complete this research, including access to research materials, data, and funding for the program. Dr. Aaron Kaulfus, who is also the lead of the CSDA program, gave me a domain knowledge perspective and highlighted why unsupervised learning is the way to go for large-scale object discovery in satellite images. I am also grateful to Iksha Gurung and Muthukumaran Ramasubramanian, computer scientists of the NASA-IMPACT team, for their continuous mentorship and support for thesis completion.

I would also like to thank my friends and family for their unwavering support and encouragement throughout this process. I would like to express my deepest gratitude to my wife, Binita Gyawali, for her love, support, and encouragement throughout the duration of this project. Similarly, I would like to thank my brother, Sudarshan Regmi, for his support and technical assistance throughout the research and writing process. His expertise and guidance were invaluable in helping me to navigate the technical aspects of this project and to produce a high-quality final product.
# Table of Contents

Abstract ................................................................. ii

Acknowledgements ....................................................... iv

Table of Contents ......................................................... vii

List of Figures ........................................................... viii

List of Tables ............................................................. ix

List of Algorithms ......................................................... x

Chapter 1. Introduction ................................................... 1

1.1 Motivation .............................................................. 1

1.2 Earth Science .......................................................... 2

1.3 Remote Sensing ......................................................... 3

1.4 Representation Learning .............................................. 3

1.5 Image Segmentation ................................................... 5

1.5.1 Semantic Segmentation ........................................... 5

1.5.2 Instance Segmentation ........................................... 5

1.5.3 Panoptic Segmentation .......................................... 6

1.6 Research Problem ..................................................... 6

1.7 Approach ............................................................... 7
Chapter 2. Background and Related Works

2.1 Instance Segmentation

2.2 Pre-training

2.3 Supervised Pre-training

2.4 Self-supervised Pre-training

2.4.1 Self-supervised Learning

2.4.2 Self-supervised Learning in Representation Learning

2.4.3 Self-supervised Learning for Dense Prediction

2.5 Weakly Supervised Learning

2.6 Self-Training

2.7 Unsupervised Learning Methods in Image Segmentation

2.8 Unsupervised Learning in Satellite Images

Chapter 3. Methodology

3.1 Free Mask

3.2 Weakly Supervised Learning in FreeSOLO

3.3 Self-Training

3.4 Optimization Loss Functions

3.5 Benchmark

3.6 Transfer Learning

3.6.1 Supervised Learning Pre-trained Weights
3.6.2 FreeSOLO Pre-trained Weights .................................. 27

Chapter 4. Results and Visualizations ................................. 30

4.1 MBRSC Dubai Aerial Imagery Dataset ................................. 31
  4.1.1 Free Mask .............................................................. 31
  4.1.2 FreeSOLO Model ...................................................... 32
  4.1.3 Comparison ............................................................. 32
4.2 PlanetScope Data (University of Alabama in Huntsville) .... 34
4.3 Benchmark Datasets (iSAID, CrowdAI, and PASTIS) ......... 36
  4.3.1 Qualitative Results .................................................. 37
  4.3.2 Quantitative Results ................................................ 38
4.4 Transfer Learning ....................................................... 38

Chapter 5. Limitations and Applications .............................. 51

5.1 Limitations ............................................................... 51
5.2 Applications ............................................................. 51

Chapter 6. Conclusion and Future Research .......................... 53

6.1 Conclusion ............................................................... 53
6.2 Future Research ........................................................ 54

References ................................................................. 56
## List of Figures

3.1 High level overview of FreeSOLO ........................................... 21
3.2 Architecture for free mask .................................................... 29

4.1 Coarse masks extracted using free mask (MBRSC dataset) .......... 33
4.2 Additional coarse masks extracted using free mask (MBRSC dataset) 42
4.3 Predicted masks using FreeSOLO model (MBRSC dataset) ......... 43
4.4 Additional predicted masks using FreeSOLO model (MBRSC dataset) 44
4.5 FreeSOLO predicted mask in PlanetScope UAH image ............ 45
4.6 Instance segmentation in iSAID satellite image ....................... 45
4.7 Additional instance segmentation in iSAID satellite image ......... 46
4.8 Class-agnostic instance segmentation in iSAID satellite image .. 46
4.9 Class-agnostic instance segmentation in a sample AICrowd satellite image ......................................................... 47
4.10 Additional class-agnostic instance segmentation in a sample AICrowd satellite image ................................................. 47
4.11 Class-agnostic instance segmentation in a sample PASTIS satellite image ............................................................... 48
4.12 Additional class-agnostic instance segmentation in a sample PASTIS satellite image ....................................................... 48
4.13 Additional class-agnostic instance segmentation in a sample PASTIS satellite image ....................................................... 48
4.14 ImageNet supervised learning based pre-trained model on down-stream segmentation task ................................................. 49
4.15 ImageNet self-supervised learning based pre-trained model on down-stream segmentation task ............................................. 50
List of Tables

4.1 Class-agnostic instance segmentation for iSAID and CrowdAI dataset 39
4.2 SL vs SSL based encodings on semantic segmentation downstream task .................................................. 40
## List of Algorithms

<table>
<thead>
<tr>
<th></th>
<th>Algorithm</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Self-Training</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>Free Mask</td>
<td>23</td>
</tr>
<tr>
<td>3</td>
<td>FreeSOLO Self-Training</td>
<td>25</td>
</tr>
</tbody>
</table>
Chapter 1. Introduction

1.1 Motivation

Machine learning is a paradigm of problem solving where the algorithm learns to solve/perform a task from the past experiences. The past experiences come in the form of data. Machine learning is basically done in two fashions – supervised and unsupervised. Unsupervised learning differs from the supervised learning in that it does not have labels or ground truth. For example, if the data contains pictures and the labels associated with the pictures (cats, dogs), the machine learning model can be developed by using those pictures and labels. This is the supervised form of machine learning. If the labels are not present, and the goal is to cluster all the pictures into different clusters, this is the unsupervised form of machine learning.

Unsupervised learning has been more important than ever. Prominent AI figures like Yann LeCun have claimed that unsupervised learning is the future of AI\(^1\). Both from the learning as well as cost viewpoint, unsupervised learning has gathered a lot of attention recently. LeCun pointed out that if machine learning, or AI, is a cake, the vast majority of the cake is self-supervised (or unsupervised) learning. The following reasons elaborate more on the motivation for doing research on unsupervised learning.

\(^1\)https://www.technologyreview.com/2019/07/12/65579/the-next-ai-revolution-will-come-from-machine-learnings-most-underrated-form/
1. **Data:** Data is the fuel that propels machine learning and deep learning algorithms. It is impossible to develop machine learning models without data because data is an integral part of model development. The model learns the patterns from data and uses the patterns to perform the prediction. For unsupervised learning, labeled data is not required. Unsupervised learning can be performed with untagged or unlabeled images. The amount of unlabeled data is five orders of magnitude higher than the amount of labeled data. So, from the perspective of data availability, unsupervised learning has an advantage over supervised learning.

2. **Cost and time:** Preparing labeled data is costly and time consuming. An example of image database having labeled images is ImageNet[9]. It has 14 million images and 22 thousand visual categories. It took roughly 22 human years to build it. The database is suited for global prediction tasks like image classification. Preparation of labeled dataset for dense prediction tasks like object detection and image segmentation is even more costly and time consuming.

1.2 **Earth Science**

It is even harder to get labeled data for specific domain like earth science because of need for domain knowledge, time to label the data, and cost associated with labeling. For example, to label earth science specific labels such as normal cloud, fire-induced cloud, or contrail, the labeler should be able to differentiate them. As is obvious, it is much harder than classifying pets as cats and dogs. In addition to that, there might be some labels which can not be visually identified. For example, the crop health status might not be distinguishable by just looking at the satellite images. Bands other than RGB might be needed to distinguish
crop health. So, additional capabilities might be needed just to label the earth science domain data.

However, unlabeled data is available at the large scale both in terms of spatial dimension as well as temporal dimension. PlanetScope\cite{48} data is offered by Planet labs, and they collect medium resolution satellite images (3m resolution per pixel) of earth approximately daily. That amounts to coverage of 200 million square kilometers per day\textsuperscript{2}. Such a large amount of data (15 TB per day) is ripe for doing unsupervised learning.

1.3 Remote Sensing

Remote sensing is a technology that allows for the acquisition of information about an object or phenomenon without physical contact. This is typically accomplished through the use of satellites or aircraft that observe the object from a distance. The information gathered through remote sensing is derived from the emitted and/or reflected radiation of the object. There are two types of sensors used in remote sensing: active and passive. Active sensors emit radiation and measure the reflected radiation to determine the properties or characteristics of the object. Passive sensors, on the other hand, only measure the reflected energy without emitting any radiation themselves. Remote sensing has proven to be a valuable tool in a variety of applications, such as land use and land cover mapping, monitoring atmospheric and environmental conditions, and more.

1.4 Representation Learning

Representation learning is the automated feature learning using some sort of machine learning paradigm (supervised or unsupervised or self-supervised). It

\footnote{https://developers.planet.com/docs/data/planetscope/}
helps to get the appropriate features from raw data to perform a particular task. Data can have many different representations, and some representations can be better than others for the end task. This fact has been exploited historically to hand-engineer important features in several domains not limited to computer vision and digital signal processing. The performance of classification systems, for instance, is largely dependent on the representation learned from the raw data [1].

The supervised form of representation learning takes both data and labels and performs the target task such as classification or regression. The features learned by the intermediate layer can be thought of as representations. The unsupervised form of representation learning does not use label information at all. Some of the examples of such representation learning are principal component analysis and matrix factorization. Self-supervised learning is the paradigm of machine learning which does not have labeled data but creates labels out of unlabeled data. It can be thought of supervised learning using unlabeled data. In this form of representation learning, features will act themselves as labels either in their pure form or modified form. Learning representations using various forms of autoencoders is an example of semi-supervised representation learning. Word2Vec, a concept of representing words in numerical embeddings, is also another example of representation learning in natural language processing.

The unsupervised or self-supervised form of representation learning has become more important than ever because of the exponential increase of unlabeled data. The first merit is being able to solve complex problems such as image segmentation in an unsupervised way. Another important merit is its powerful capability to do representation learning without the use of labels. Unsupervised learning or self-supervised learning techniques help learn better representations of the data, and this can improve the performance of downstream tasks. The third
merit of unsupervised/self-supervised learning is bettering the semi-supervised learning based-systems using unsupervised/self-supervised learning components or embeddings returned by the unsupervised/self-supervised learning components.

1.5 Image Segmentation

Image segmentation is the pixelwise classification of an image into several categories. If an image has height $h$, width $w$, $c$ channels, and $k$ number of categories, then image segmentation transforms the image from shape $h \times w \times c$ to $h \times w \times k$. Image segmentation has further three groups. They are semantic segmentation, instance segmentation, and panoptic segmentation. They are discussed in more detail in the following subsections.

1.5.1 Semantic Segmentation

It is the group of image segmentation where the image is segmented into different semantic categories. For example, if an image has two people and a background, two people are segmented as one class i.e., people and the background is segmented as another class.

1.5.2 Instance Segmentation

Instance segmentation is similar to semantic segmentation but it segments instances of the same class separately. In the example above, the two people are segmented differently. So, there are three instances segmented in total i.e., first people, second people, and background.
1.5.3 Panoptic Segmentation

Panoptic segmentation is the combination of semantic segmentation and instance segmentation. It segments an image into semantic categories, yet also differentiates different instances of the semantic categories. So, in the above example, there is semantic segmentation of the image into two categories i.e. people and background. There is also further segmentation of the two people into two instances of people. This segmentation is more “complete” compared to other two groups of segmentation.

Instances can be both things and regions of stuff. *Thing* refers to countable objects. For example, car, house, and person can be considered as *thing* as it can be counted. *Stuff* (or region of stuff) refers to amorphous object which can not be counted. For example, the region of road or sky can not be counted. So, it refers to *stuff* (or region of stuff).

1.6 Research Problem

Unlabeled images are everywhere. The high-frequency medium resolution PlanetScope[48] data was available through the Commercial Smallsat Data Acquisition (CSDA)[34] Program. Supervised learning was not possible in the PlanetScope data because labels were not available. So, the first research problem was to segment the images into different instances in an unsupervised fashion. Here, the instance can be either thing or stuff. For satellite images, sea or grassland can be stuff. On the other hand, ship or house can be thing. Apparently, the segmentation of satellite images into different objects has a lot of applications in satellite image search and object discovery. As domain-agnostic unsupervised instance segmentation was not found for satellite images in the literature, the
second research problem was to establish a benchmark for it. Benchmarking was considered essential because researchers could perform a comparative analysis of their methods in the future. The third research problem was the comparison of the representation learned by the FreeSOLO-based encoder with the supervised learning-based encoders based on dense prediction downstream tasks. If a small number of labeled images are available for some downstream tasks, can the FreeSOLO-based representations be used for those downstream tasks? How do those representations compare against supervised learning-based representations? These are the questions the research aims to answer.

1.7 Approach

The approach to solving the research problem is to make use of self-supervised learning to segment images in an unsupervised way. This research uses contrastive learning under the self-supervised learning paradigm. At first, the pretrained model trained in a self-supervised way using contrastive learning is taken. The novel idea of using dense contrastive learning for building self-supervised pre-trained model was introduced by [54]. Their idea as well as their pretrained model is used for this research. They have also introduced free mask[53] concept to get course mask of the instances from the image using the dense contrastive learning based pre-trained embedding model. Their free mask approach is replicated to get coarse masks for satellite data. [53] use the coarse mask to do weakly supervised learning using BoxInst approach[51]. Using these two concepts, free mask and self-supervised learning, they have build a model to segment objects in an unsupervised way. The final instance segmentation model is used to segment the objects in the satellite images. The segmentations of the
initial free mask[53] approach are also compared with the final FreeSOLO[53] model in the satellite images.

The iSAID[56], CrowdAI[37], and PASTIS[44] datasets are used to establish the satellite imaging-based domain agnostic unsupervised instance segmentation baseline. The FreeSOLO model is run on these three datasets and the baseline is established.

To assess and answer the question about the effectiveness of the unsupervised learning-based pre-trained weights, their performance is compared with supervised learning-based pre-trained weights in downstream semantic segmentation tasks with respect to different segmentation architectures.

1.8 Thesis Organization

The thesis is organized into six chapters. The first chapter introduces readers to the thesis topic and research problem. It also gives the motivation behind the research problem and the approach to solving it. The second chapter gives more background to the research topic and provides related works. It gives a brief literature review on unsupervised image segmentation and unsupervised techniques used in satellite images. It also presents the various unsupervised learning methodologies in the existing literature to segment satellite images. The third chapter explains the methodology used for this research work in detail. This research work is application of computer science knowledge in earth science domain (i.e. satellite images). The computer science knowledge (i.e. self-supervised image segmentation) is explained in great detail in this section. The fourth chapter presents the result of the experiments. First, the self-supervised image segmentation method is used in labeled dataset to compare the result with ground truths. The model is benchmarked on three remote sensing instance seg-
mentation datasets. Then, it is used in the University of Alabama in Huntsville periphery to do visual inspection using PlanetScope data. It is also used on MBRSC Dubai dataset to see the instance segmentation performance. Following these experiments, the self-supervised learning based pre-trained weights are used in semantic segmentation downstream task. The performance of the self-supervised learning based pre-trained weights is compared with the performance of supervised learning based pre-trained weights. The comparison is done with respect to semantic segmentation downstream task. The fifth chapter discusses the limitations of this method and the origin of its limitations. The sixth chapter provides the conclusion of the work and its possible future directions.
2.1 Instance Segmentation

Instance segmentation is segmentation of an image into different instances of objects. Top-down methods of instance segmentation first detect bounding box of the objects and then segment the objects within the bounding boxes[27][16][4][31]. Bottom-up methods learn embedding for each pixel in an image and then group the pixels into different segments[8][39][12]. There are also some methods which combine top-down and bottom-up methods[2][50]. All in all, there have been a lot of progress in instance segmentation using supervised learning[55][50][4]. However, supervised instance segmentation methods require rich annotations of the training data, which is both difficult and costly to obtain. Various weak supervised methods such as BoxInst[52] and DiscoBox[25] have been proposed for instance segmentation tasks. They have reduced the performance gap between supervised and weakly supervised methods. Nonetheless, they still require some kind of annotations such as bounding box or point information. State-of-the-art unsupervised instance segmentation method, FreeSOLO[53], has shown fully unsupervised instance segmentation for the first time.

2.2 Pre-training

Pre-training is a popular technique in machine learning where a large model is trained using large amount of data and the weights of all the layers are saved so that they can be reused. The weights can be reused when solving
a similar problem with less data. The last few layers can be cut off in the pre-trained model and the custom layers can be added based on the usecase. Then, the model can be finetuned using custom data. Pre-training is pretty much common in computer vision. Convolutional neural networks are seen to learn hierarchical features[26]. Low level features such as edges and shapes are common in almost all the computer vision tasks. So, the layers act as feature extractors and they help to get better representation of the raw pixel data.

2.3 Supervised Pre-training

ImageNet pre-trained classification model is one of the pre-trained model which has been used since years. The pre-trained model is developed using ImageNet dataset for image classification task. This form of pre-training that uses labeled data is called supervised pre-trained model. The pre-trained model performs well for image classification task but there is performance gap when it is used to perform dense prediction tasks such as image segmentation and object detection[15]. It is because of the difference between the nature of the tasks – image classification assigns a category to a whole image whereas image segmentation assigns category to each pixel. The straightforward approach to solve this problem is to develop a pre-trained model on dense prediction tasks themselves. However, this is difficult to do as it is hard to label data for dense prediction tasks. It is not as straightforward as assigning labels to each image. So, one way to approach this problem is to go with self-supervised pre-training method.
2.4 Self-supervised Pre-training

2.4.1 Self-supervised Learning

In self-supervised learning, the explicit label information is not available as part of training data. It builds the labels from the features themselves in a clever way. For example, an image is given some noise and that gives a pair of noisy image and the original image. If the noisy image is supposed as the input and the original image as the output, it gives denoising autoencoder architecture[59]. This is an example of self-supervised learning. In this example of self-supervised learning, the objective function has a reconstruction loss component in it. Reconstruction-based loss functions are one class of objective functions common within self-supervised pre-training[11][13].

Self-supervised learning can also have contrastive loss as the objective function. In this type of self-supervised learning, two types of pairs are formed - positive pair and negative pair. The two different views of the same image can be a positive pair and the pair of two different images can be a negative pair. The training examples of positive pair are pulled together and the training examples of negative pair are pushed apart. In this way, feature representation of the training data is learned. So, contrastive loss function is also another common loss function used with self-supervised pre-training[49].

2.4.2 Self-supervised Learning in Representation Learning

SimCLR[5] and MoCo-v1/2[14][6] are the breakthrough approaches of self-supervised representation learning. Basically, a set of unlabeled images are given and the better representation is learned by learning for instance discrimination[57] pretext task. In instance discrimination pretext task, the positive pair has differ-
ent views of the same image whereas negative pair has views of different images. The views are different instances of the same image formed by applying some transformations to the image. In this type of representation learning, the features of an image are pushed apart to the features of different images.

While doing pre-training, there are two parts – backbone and projection head. The backbone is taken from some standard architecture such as ResNet[17] and the projection head is a stack of layers stacked on top of the backbone. Two views of same or different images are passed through backbone and projection head. The result will be global feature vector. One will be query feature vector and another will be key feature vector. For each query feature vector $q$, there is a key feature vector $k$ that matches with the query. Such a pair of query and key are from the positive pair. The InfoNCE loss[41] is used as the objective function such that positive pairs are pulled together and the negative pairs are pushed apart.

$$L_q = - \log \left( \frac{e^{q, k_+}}{e^{q, k_+} + \sum_{k_-} e^{q, k_-}} \right)$$  \hspace{1cm} (2.1)$$

where $\tau$ is temperature hyperparameter.

Some of the self-supervised pre-training methods have shown comparable or better performance than supervised pre-training methods on image classification tasks[5][6][14]. However, there is still gap between pre-training using image classification task and downstream dense prediction tasks. So, DenseCL[54] has been proposed to further improve the performance of downstream dense prediction tasks. DenseCL[54] uses dense projection head instead of global projection head so that the spatial representation of features is preserved.
2.4.3 Self-supervised Learning for Dense Prediction

In DenseCL[54] architecture, the backbone is connected to two parallel heads i.e. global projection head and dense projection head. Global projection head is the same as in the standard self-supervised pre-training architectures described above. It outputs the global feature vector. The dense projection head is similar to global projection head but it does not do global pooling and replaces the multilayer perceptron layer with $1 \times 1$ convolutional layer.

The feature maps produced by the dense projection head will have the height of $H_f$ and width of $W_f$. Here, a query $r$ does not represent the whole image but a part of the image. For a query $r$, there are a number of keys $t_0, t_1, ...$. The negative key, $t_-$, for the query is the pooled feature vector of a view of different image. The positive key, $t_+$, is the correspondence key of the different view of the same image. So, the contrastive loss coming from the dense projection head will be as follows:

$$L_r = -\frac{1}{H_f W_f} \sum_i \log \left( \frac{e^{r_i t'_+}}{e^{r_i t'_+} + \sum_t e^{r_i t'}} \right)$$

(2.2)

where $\tau$ is temperature hyperparameter and $i$ belongs to one of the $H_f \times W_f$ indices.

Now, the DenseCL[54] uses total loss as the weighted combination of (2.1) and (2.2).

$$L = (1 - \lambda)L_q + \lambda L_r$$

They have set $\lambda$ to 0.5 as validated by their experiments.
2.5 Weakly Supervised Learning

Preparing training data can be the biggest hurdle in developing machine learning-based systems, especially when the training data needs to be labeled. So, many other techniques that make use of very less to no labels are rapidly developing. As a result, various machine learning paradigms other than supervised learning are evolving rapidly.

Weakly supervised learning is the paradigm of machine learning which attempts to use both label information and unlabeled images together to improve the performance of the fully supervised or fully unsupervised methods. Weakly supervised learning methods (using \( D_1 \) labeled data and \( D_2 \) unlabeled data) are aimed to improve the performance of both supervised (using just \( D_1 \) data) and unsupervised (using unlabeled data, both \( D_1 \) and \( D_2 \)) learning-based systems.

Weakly supervised learning is broadly classified into three types[62]:

- **Incomplete supervision**: This is the type of weakly supervised learning where a subset of the training data has ground truth. Out of the whole training data \( D \), the subset of training data \( D_1 \) has ground truth, and the subset of training data \( D_2 \) does not. Normally, the size of \( D_2 \) is way bigger than the size of \( D_1 \). The aim of this type of weakly supervised learning method is to develop a model which performs better than the model developed by either unlabeled data from \( D \) or labeled data \( D_1 \).

- **Inexact supervision**: In this type of weakly supervised learning, the ground truth is present for all the training examples in the training data but they are not exact; they are coarse. For example, the exact ground truth information for the image segmentation task is contained in the im-
age segmentation masks. Bounding boxes, on the other hand, will act as the coarse ground truth for the image segmentation tasks.

- **Inaccurate supervision:** Similar to inexact supervision, the ground truth information is present for all the training examples in the form of labels. This method contrasts with the inexact supervision method in that the labels may not always represent ground truth or be accurate. The inaccuracy of the labels may come from incorrect labeling. One example where this type of method can be used is with crowdsourcing data[3]. Because of high cost of collecting labeled training data, crowdsourcing has been used. Such a technique of data collection can give rise to inaccuracies. This type of method can be used with such type of data.

It is worth mentioning that weakly supervised data do not come in the purest form as far as the supervision types are concerned. Rather, two or more types come simultaneously in the data. So, the weakly supervised learning methods are designed to work with multiple types of weakly supervised data.

**2.6 Self-Training**

One type of weakly supervised data is incomplete supervised data. Incomplete supervised data have labels or ground truths for just a subset of the whole data. One technique used to train a predictive model using this type of data is self-training[45]. This technique has a network called pseudo-labeler ($N_{pl}$) that labels the unlabeled data. Then, the labeled data is concatenated with pseudo-labeled data, and the model is trained from the scratch using the concatenated data. One simple example of a pseudo-labeler can be a predictive model developed using the labeled data.
The labeled data is denoted as $D_l$, which is just a set of features and labels i.e. $\{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$. Similarly, the unlabeled data is denoted by $D_u$, which is a set of features i.e. $\{x_{l+1}, \ldots, x_N\}$. $N$ denotes the total number of samples for both labeled and unlabeled data. Now, the classic self-training algorithm is presented below as Algorithm 1.

**Algorithm 1 Self-Training**

1: Train a pseudo-labeler network, $N_{pl}(.)$, using the labeled data, $D_l$.
2: repeat
3: Use $N_{pl}$ to pseudo-label the unlabeled data, $D_u$.
4: Subset pseudo-label data into $S$ such that $x \in D_u$ and $(x, N_{pl}(x)) \in S$.
5: Remove $S$ from $D_u$.
6: Train the predictive network, $N_{pl}$, using data $D = D_l \cup S$.
7: until convergence or $D_u = \phi$ or the iterations are over.

### 2.7 Unsupervised Learning Methods in Image Segmentation

One common approach to image segmentation is clustering. Image segmentation is basically the partitioning of an image into multiple segments based on pixel information. Similar groups of pixels are clustered together to form an image segment based on the characteristics like color, intensity, contrast, and semantic meaning. K-means clustering is the most popular clustering algorithm which has been used across different fields to do image segmentation[22][10].

K-means algorithm and fuzzy C-Means algorithm have been used in the field of medicine to segment brain tumors and assist in their area calculation[22]. They first preprocess the images to improve their quality. Then, they use their proposed K-means and fuzzy C-means method to do segmentation. Then, they extract features from the image segments. There has also been a lot of other research which uses k-means on its own or combined with other methods in the field of medicine[40][23][38].
Spectral clustering can also be used to perform image segmentation and has been widely used\cite{46}\cite{7}\cite{61}. It is a method for clustering data that is based on the eigenvalues and eigenvectors of a matrix derived from the data. At first, a similarity matrix is constructed using the pairwise distance between the data points. Then, the Laplacian matrix is constructed. The Laplacian matrix encodes the connectivity of the data points. The Laplacian matrix is typically constructed from the similarity matrix. After that step, eigenvectors of the Laplacian matrix are calculated, and the data points are clustered in the low-dimensional space. Finally, the clusters are mapped back to the original data space. That gives different clusters corresponding to the different segments of the image.

Deep learning has also been used for unsupervised image segmentation. W-net\cite{58} uses the encoder-decoder architecture of the convolutional neural networks to segment images in an unsupervised way. Other methods involving convolutional neural networks and backpropagation have also been studied showing promising results\cite{24}.

Melas et al.\cite{35} propose a deep spectral approach for unsupervised semantic segmentation and localization. Specifically, the authors construct a Laplacian matrix from a combination of color information and deep features obtained in an unsupervised manner. The proposed method demonstrates superior performance in comparison to the state-of-the-art for unsupervised image segmentation and object localization.

Inspired by autoregressive generative models, \cite{42} have proposed an unsupervised image segmentation method based on the maximization of mutual information between different views of the same image.
2.8 Unsupervised Learning in Satellite Images

Tile2Vec[21], a representation learning method for spatially distributed data, was proposed to generate lower-dimensional embeddings for higher-dimensional tiles. They used the concept of spatial neighborhood in determining similar tiles and different tiles. Bringing closer the embeddings of similar tiles and pushing farther the embeddings of different tiles, they were able to get better representation of the tiles using triplet loss based objective function. They were able to outperform other unsupervised feature extraction techniques and also showed the better performance of their embeddings in downstream tasks such as land cover classification and poverty prediction.

Self-supervised learning has been explored as a pre-training method for learning representation of satellite images[33]. Seasonal Contrast (ScCo)[33] was proposed to fill the domain gap of using an ImageNet[9]-based pre-trained model in satellite imagery-based earth science prediction tasks. They presented the automated data acquisition pipeline to acquire satellite imagery data. They also presented the self-supervised pre-training architecture using spatiotemporal satellite data. They used the temporal data as the natural image augmentation, and applied them in conjunction with artificial image augmentation to do self-supervised contrastive learning. They were also able to outperform the state-of-the-art methods for land-cover classification and change detection when their pre-trained model was applied to the downstream tasks.
Chapter 3. Methodology

In this chapter, the methodology used to perform instance segmentation in satellite images is described. As FreeSOLO[53] has been used to segment instances in satellite images, the methodology of FreeSOLO is thoroughly explained and the usage of the methods and algorithms of FreeSOLO in the research are described. FreeSOLO[53] uses DenseCL[54] pre-trained model and as such, DenseCL[54] pre-trained model has been used. The pre-trained model is used to extract coarse masks. The algorithm to extract coarse masks is presented as it is used in FreeSOLO[53]. Then, all the procedures to further train the instance segmenter are described taking reference to the original FreeSOLO work and its preliminaries. Figure 3.1 shows the high-level overview of the FreeSOLO[53] method.

Also, the details about how the baseline is established for satellite imaging-based class-agnostic unsupervised instance segmentation tasks are described. Finally, the experiments run with respect to the semantic segmentation downstream task are described. One set of experiments takes supervised learning-based ResNet-101 backbone pre-trained weights. Another set of experiments takes self-supervised learning-based ResNet-101 backbone pre-trained weights. Then, the performance of supervised learning-based weights and self-supervised learning-based weights is compared to each other with the help of the metrics, intersection over union (IOU), and dice coefficient.
3.1 Free Mask

The first step in segmenting the instances in an image is extracting the coarse masks of the instances. The coarse masks are not perfect masks but pseudo-masks that spatially localize the objects. The free mask is the method of generating coarse masks of the instances by using the key-query mechanism. It is introduced in FreeSOLO[53] as the first step to segment objects in an unsupervised way. The free mask described in their paper uses pre-trained backbone trained in a self-supervised way. They use the pre-trained backbone trained using dense contrastive learning, DenseCL[54], to generate the free mask. The unlabeled image is passed through the pre-trained backbone and a set of feature maps of shape $H \times W \times E$ are produced. The feature maps are bilinearly downsampled to produce queries, $Q \in R^{H' \times W' \times E}$. The feature maps themselves act as the keys, $K \in R^{H \times W \times E}$. Now, the score maps ($S$) are calculated by finding cosine similarity of each query with all the keys giving score maps, $S \in R^{H \times W \times H' \times W'}$. The operation can be written as:

$$S_{i,j,q} = cosim(Q_q, K_{i,j})$$  \hspace{1cm} (3.1)

where $cosim$ is the cosine similarity, $Q_q \in R^E$ is the $q$th query, and $K_{i,j} \in R^E$ is the key in position $(i,j)$. The cosine similarity of the two vectors $\vec{x}$ and $\vec{y}$ is
given by the dot product of \( \vec{x} \) and \( \vec{y} \) divided by product of their L2-norms. So,

\[
\text{cosim}(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|}.
\]

Then, the score maps are passed through min-max normalization to form soft masks. Soft masks have values in the range of \([0, 1]\). Now, soft masks are applied some threshold, \( \tau \), and the binary masks are formed. All of the soft masks have maskness score. Maskness score is the non-parametric score used to rank the coarse masks. Maskness score of a soft mask is given by the following formula.

\[
\text{maskness} = \frac{1}{N_f} \sum_{i}^{N_f} p_i
\]  

(3.2)

where \( N_f \) is the number of pixels which have the normalized value greater than the threshold, \( \tau \), and \( p_i \) is the normalized value at the \( i \)th pixel.

Then, the masks are sorted on the basis of maskness score of the masks, and non-maximum-suppression (NMS) is used to remove the redundant masks. After redundant masks are removed, the coarse masks are obtained. The architecture diagram for this component is shown in figure 3.2.

The algorithm can be written as given in Algorithm 2.

### 3.2 Weakly Supervised Learning in FreeSOLO

It has already been mentioned in the free mask section how FreeSOLO\cite{53} has successfully used key-query mechanism in the self-supervised pretrained model to extract coarse masks. The coarse masks are the weak labels. These type of masks can be incomplete supervision as they might not have coarse masks for all the objects. They can also be inexact supervision because they might not segment the exact shape of the objects. Also, they can be inaccurate supervision as they may extract inaccurate masks of the objects.
Algorithm 2 Free Mask

1: Use DenseCL pre-trained backbone to generate feature maps of shape $H \times W \times E$ from images of shape $H \times W \times C$.
2: Duplicate feature maps to produce keys of shape $H \times W \times E$.
3: Perform bilinear downsampling to produce queries of smaller shape $H' \times W' \times E$.
4: Calculate cosine similarity between keys and queries to produce score maps of shape $H \times W \times H' \times W'$.
5: Use min-max normalization to normalize the score maps to the range $[0, 1]$. They are termed soft masks. Each soft mask has its maskness score.
6: Use some threshold to produce binary masks from soft masks.
7: Sort the binary masks using maskness score.
8: Use NMS to further filter the binary masks and final coarse masks are obtained.

Dice loss[36] is used in SOLO[55] to learn the instance segmenter using the ground truth object masks. FreeSOLO[53] uses the coarse masks and the semantic embeddings, both obtained from the free mask extractor, to learn the SOLO-based instance segmenter. However, the coarse masks are not directly used to supervise the ground truth masks as they may be inaccurate and inexact, and they may not give good results. So, weakly supervised learning is used to learn the instance segmenter using the extracted free masks and semantic embeddings.

BoxInst[52] is a weakly supervised learning method to segment objects using bounding box notations. They project the bounding boxes in $x$-axis and $y$-axis, and use these projected vectors in the loss function. FreeSOLO[53] also uses the similar approach to perform weakly supervised learning. The weakly supervised learning approach they have used is explained below.

Let $m$ and $m^*$ denote the predicted and coarse mask for an object. Then, the predicted and coarse mask are projected on the $x$-axis and $y$-axis using $max$ as well as $avg$ projection. So, the two losses, $max$ projection loss and $avg$ projection loss, are written mathematically as given below.
For max projection loss,

\[ L_{\text{max proj}} = L(\max_x(m), \max_x(m^*)) + L(\max_y(m), \max_y(m^*)) \]  \hspace{1cm} (3.3)

For avg projection loss,

\[ L_{\text{avg proj}} = L(\text{avg}_x(m), \text{avg}_x(m^*)) + L(\text{avg}_y(m), \text{avg}_y(m^*)) \]  \hspace{1cm} (3.4)

where,

\( L(., .) \) represents Dice loss[36],

\( \max_x \) represents maximum projection along \( x \)-axis,

\( \text{avg}_x \) represents average projection along \( x \)-axis,

\( \max_y \) represents maximum projection along \( y \)-axis, and

\( \text{avg}_y \) represents average projection along \( y \)-axis.

They also have another loss component called pairwise affinity loss[52],
which tends to put neighboring similar valued raw pixels in the same instance.
So, the total loss is defined as:

\[ L_{\text{mask}} = \alpha L_{\text{avg proj}} + L_{\text{max proj}} + L_{\text{pairwise}} \]  \hspace{1cm} (3.5)

where \( \alpha \) is the hyperparameter to balance weight of different loss components.

### 3.3 Self-Training

FreeSOLO[53] uses coarse masks obtained from a free mask extractor to do weakly supervised learning. The weakly supervised learning is done in the SOLO-based architecture. The resultant masks obtained from the SOLO-based segmenter after doing weakly supervised learning are better (both qualitatively
and quantitatively) than the free masks. As part of self-training, the top predicted masks (predicted by SOLO-based segmenter, not free mask extractor) are sent to the weakly supervised algorithm again as weak labels, and the SOLO-based segmenter is trained. The performance of the SOLO-based segmenter increases using this approach of self-training. In this way, they make use of self-training with SOLO-based architecture.

The free mask extractor is denoted as $M_c$, which is used in the first step to extract coarse masks from an unlabeled image. Using coarse masks and SOLO-based architecture, the pseudo-labeler network ($N_{pl}$) is trained, which uses weakly supervised learning. The masks outputted by the pseudo-labeler network are symbolized as $M_i$, and the top instance masks subset is symbolized as $M_{it}$. The algorithm for FreeSOLO self-training is presented below as Algorithm 3. It repeats until the performance does not increase (i.e., convergence) or the absence of top instance masks or the end of a fixed number of iterations.

**Algorithm 3** FreeSOLO Self-Training

1: Extract coarse masks ($M_c$) using free mask extractor.
2: Train pseudo-labeler network ($N_{pl}$) using coarse labels and SOLO-based segmenter using weakly supervised learning.
3: repeat
4: Use $N_{pl}$ to extract instance masks, $M_i$.
5: Subset top instance masks based on their confidence scores.
6: Train $N_{pl}$ in weakly supervised fashion using top instance masks, $M_{it} \subseteq M_i$.
7: until convergence or $M_{it} = \emptyset$ or the iterations are over.

### 3.4 Optimization Loss Functions

As in SOLO[55] architecture, the optimization loss function has two loss components. One is mask loss ($L_{mask}$, as defined by equation 2.7) and another is categorical loss ($L_{cat}$). The categorical loss also contains two further components.
The architecture is different than that of SOLO because this architecture has two components in the category head. The first component is standard Focal loss [29]. The second component is categorical semantic loss, which learns the output of the embedding by the free mask. If the output of the embedding given by the free mask is \( q^* \) and the predicted output of the embedding is \( q \), the negative cosine similarity loss is given by:

\[
L_{\text{sem}} = 1 - \frac{q}{\|q\|_2} \frac{q^*}{\|q^*\|_2}.
\]  

(3.6)

This loss is added to the focal loss, so, the total category loss is given by:

\[
L_{\text{cat}} = L_{\text{focal}} + \lambda L_{\text{sem}}.
\]  

(3.7)

Now, the total loss of the whole architecture becomes the sum of the loss from the category branch and mask branch. The expression for the loss of the mask branch is given by equation 2.7 whereas the loss for the category branch is given by equation 2.9.

### 3.5 Benchmark

In this study, three distinct satellite imagery datasets are utilized, namely iSAID[56], CrowdAI[37], and PASTIS[44], to benchmark the efficacy of the FreeSOLO[53] method in the context of class-agnostic unsupervised instance segmentation in satellite images. Due to the significant size of images within the iSAID dataset, the images are partitioned into 256 pixels by 256 pixels and the experiments are conducted in them. Both the lower-size data and lower-size annotations had to be created – still aligning with the COCO format – from the dataset. The bench-
mark was established based on the outcomes obtained from the 256 x 256-sized images and the corresponding annotations. The PASTIS dataset, on the other hand, was not in COCO format. So, data preprocessing was done to bring the images and annotations to the COCO format.

3.6 Transfer Learning

Several experiments were conducted to evaluate the performance of pre-trained weights on semantic segmentation downstream tasks. The standard metrics for semantic segmentation, intersection over union (IOU) and dice coefficient, are used as an evaluation or comparison metric.

3.6.1 Supervised Learning Pre-trained Weights

Initially, the backbones corresponding to the ResNets, ResNeXts, ResNeSt, RegNet, GERNet, SE-Net, DenseNet, and VGG encoders are employed. Each of these backbones is utilized individually to extract features from the RGB satellite images, thereby serving as the encoder for generating image embeddings. Subsequently, the weights and biases of each backbone are held constant, and the layers from various segmentation architectures are assembled atop each backbone. This framework forms the basis for the semantic segmentation architecture, which is trained on the labeled images of the training dataset and evaluated on the held-out test data. Notably, each backbone model is trained and assessed independently.

3.6.2 FreeSOLO Pre-trained Weights

The present study utilizes pre-trained weights sourced from the ResNet-101 backbone of the FreeSOLO model, which were trained through the FreeSOLO method of self-supervised learning. These weights serve as an encoder, which is
integrated with a semantic segmentation architecture. To investigate the efficacy of this approach, experiments were conducted using both the feature pyramid network and U-Net architecture with the FreeSOLO ResNet-101 backbone. The encoder weights were maintained as unmodifiable, while the decoder network was trained in a supervised manner, using the same training data. The resulting transfer learning methodology employed the self-supervised learning-based pre-trained weights of the backbone for the downstream task of semantic segmentation. Finally, the model was evaluated using held-out test data and compared against supervised learning-based embeddings.
Figure 3.2: Free mask[53] architecture
Chapter 4. Results and Visualizations

In this chapter, the results of the unsupervised instance segmentation work are shown using FreeSOLO[53] in many different datasets. The qualitative visualization is shown in MBRSC Dubai Aerial Imagery dataset[20] and PlanetScope satellite imagery data[48] as they don’t have instance labels. The former dataset is a standard labeled dataset for semantic segmentation tasks. The PlanetScope data is the unlabeled data obtained through Commercial Smallsat Data Acquisition Program (CSDA) program[34]. The former data is used to see and evaluate the segmentation results with the ground truth masks. The latter data is used to actually test the algorithmic procedure on unsupervised images.

In addition, the qualitative and quantitative results are shown on three satellite image instance segmentation datasets – iSAID[56], CrowdAI[37], and PASTIS[44]. iSAID is the first benchmark dataset for instance segmentation in aerial images. CrowdAI is the instance segmentation dataset for buildings that initially appeared as an AI Crowd mapping challenge. PASTIS is a semantic and panoptic segmentation dataset of agricultural parcels containing Sentinel-2 multispectral images.

The FreeSOLO[53] segmentation procedure has two models, free mask extractor and final pseudo mask extractor. Both of these models (or procedures) are used on MBRSC Dubai dataset to visualize and compare the coarse mask and the final masks.
4.1 MBRSC Dubai Aerial Imagery Dataset

The dataset is published by Humans in the Loop\(^1\) on Kaggle\(^2\) in collaboration with Mohammad Bin Rashid Space Center\(^3\). The dataset is published for open access, so the data was accessed through Kaggle. The dataset contains aerial imagery of Dubai obtained through MBRSC satellites. It has image tiles and the corresponding segmentation masks having six classes – buildings, land, road, vegetation, water, and unlabeled. The data was segmented by trainees of Roia Foundation in Syria.

4.1.1 Free Mask

All the satellite images of MBRSC Dubai Aerial Imagery Dataset are run through the free mask extractor and the sets of free masks are produced. The free masks give coarse masks of the images pertaining to different instances. It is to be noted that different instances of the objects are segmented here instead of the different semantic categories. The visualization presented below might give the misconception that the segmenter is segmenting the semantic classes. It is not segmenting the different semantic classes. It is segmenting different instances of the objects, agnostic to the semantic classes. However, the visualization is prepared in such a way that similar semantic instances are given the same color for the convenience of the reader and better interpretability.

Figure 4.1 shows the qualitative results of free mask\(^{53}\) on the MBRSC Dubai Aerial Imagery dataset. As can be seen, the water bodies (lakes and rivers) and settlement areas are segmented as different instances. From the view

\(^1\)https://www.kaggle.com/datasets/humansintheloop/semantic-segmentation-of-aerial-imagery
\(^2\)https://www.kaggle.com/
\(^3\)https://www.mbrsc.ae/
of satellites, the regions of stuff such as lakes and settlements appear to be like instances of everyday objects. So, the technique of instance segmentation can be extended to segment regions of stuff for satellite images.

Figure 4.2 shows more visualizations of free mask output. It is fascinating to see how this approach can segment complex shapes of rivers too (as demonstrated in the second row of Figure 4.2). One thing worth noticing is that this approach does not segment coarse masks as well for big regions of stuff as it does for small regions of stuff. This makes some sense as the DenseCL\[54\] pretrained model was trained on everyday objects. Everyday objects are things rather than regions of stuff. Even if everyday objects may have some regions of stuff, they are not as frequent as in satellite images.

### 4.1.2 FreeSOLO Model

The satellite images of MBRSC Dubai Aerial Imagery Dataset are then run through FreeSOLO\[53\] model. The model performs better instance segmentations than free mask as seen in the figure 4.1 and figure 4.2. The same images used to visualize free mask\[53\] results are used to visualize the FreeSOLO\[53\] model results for visual comparative analysis.

### 4.1.3 Comparison

Segmentation results of free mask and FreeSOLO are compared in this subsection. The visualization of the segmentation methods show some similarity, yet are different in some respects. Overall, it can be seen that FreeSOLO model segments the instances better than the free mask. The comparison between them is done with respect to following points.
1. **Completeness:** Comparing free mask[53] and FreeSOLO[53] model results with respect to figure 4.1 and figure 4.3, it can be seen that FreeSOLO model results are more complete than free mask results. For example, in the top image of figure 4.1, free mask segments just the water bodies whereas FreeSOLO model segments land area, building area, and little bit of road area too in addition to the water bodies.

2. **Shape:** The shape of the instances are well taken care of by FreeSOLO model compared to free mask. It makes sense as the purpose of free mask is
to generate coarse masks to do weakly supervised learning. This is validated by the visualization results too.

3. **Coverage:** The area coverage is wider for FreeSOLO segments in comparison to free mask segments. The FreeSOLO segments have higher recall but lower precision whereas free mask segments have higher precision but low recall. It is also because FreeSOLO gives a lot of segmentation masks compared to free mask.

4.2 **PlanetScope Data (University of Alabama in Huntsville)**

PlanetScope\[48\] data was available through Commercial Smallsat Data Acquisition (CSDA) program\[34\]. The satellite image of the University of Alabama in Huntsville (UAH) area is taken through planet data explorer. For inference, “Visual” tiles are taken. “Visual” tiles are targeted for visual analysis. Those tiles contained three channels i.e. red, green, and blue. It was fit for the FreeSOLO\[53\] model as the model was trained in COCO dataset\[30\] and the dataset had three visual channels i.e. red, green, and blue.

The PlanetScope data had resolution of 3m per pixel. The medium resolution image was subsetted into small images by taking five rows and three columns. Then, a total of fifteen images were passed through the FreeSOLO\[53\] model. The model gave instance segments in terms of segmentation masks. The segmentation masks were given different colors manually with respect to its semantic category. Then, all the results were merged to form of single big image. Finally, the original image and segmentation results were put side by side to do visual comparison as shown in figure 4.5.
As shown in figure 4.5, several meaningful objects were segmented. The model performed the best in segmenting the UAH university lake. It segmented the whole lake. As also is the case with Dubai MBRSC dataset, the model seems to perform better in segmenting water bodies like lakes and rivers. The lake segmentation produced by the model is given a blue mask.

In addition to the lake, the model also performs well in segmenting the visually identifiable buildings, especially the ones having rectangular shape. The buildings are given mask color of light blue. The following buildings were segmented by the model.

1. Lookheed Martin Space
2. 4800 Bradford Dr NW
3. 408 Allen St NW
4. Alabama Technology Network/National Weather Service
5. Teledyne Brown Engineering
6. Olin B. King Technology Hall
7. Shelbie King Hall
8. Johnson Research Center
9. University Fitness Center
10. Charger Village

The third class of objects that the model was able to segment was housing area. The housing areas are given the mask color of cyan. It can be seen that the model is segmenting the fraternity row, the Southeast Campus Housing, and
the housing areas around the back entrance of UAH. It is also interesting to see
the model segmenting some portion of I-565 highway. The mask color for roads
and highway is given yellow. In addition to them, the model is segmenting the
parking lot of Teledyne Brown Engineering and Charger Park, and two green
areas covered with trees. The parking lot is given reddish mask color and the
treed area is given green mask color.

4.3 Benchmark Datasets (iSAID, CrowdAI, and PASTIS)

iSAID dataset is the first satellite image instance segmentation dataset.
The dataset contains high spatial resolution images containing class and instance
annotations of fifteen classes of images i.e. plane, ship, storage tank, baseball
diamond, tennis court, basketball court, ground track field, harbor, bridge, large
vehicle, small vehicle, helicopter, roundabout, soccer ball field, and swimming
pool.

CrowdAI dataset is the instance segmentation and semantic segmentation
dataset which initially appeared as a machine learning modeling challenge. It has
instances of buildings in its dataset. It also contains instance annotations. The
dataset has medium-resolution spatial images.

PASTIS is the dataset used as a standard for the classification of agricul-
tural land using satellite images. It has both instance index and semantic labeling
for each pixel. The use of the instance index is made for instance segmentation.
Each of these patches is made up of a series of Sentinel-2 multispectral images
that vary in length, with a size of 125 pixels by 125 pixels. RGB bands are used
to do the instance segmentation. The agricultural parcels have been categorized
into 18 different crop types. Additionally, there is a background class for non-
agricultural land and a void label for areas that are primarily located outside the patch.

4.3.1 Qualitative Results

The two figures 4.6 and 4.7 show the visualization of the instance segmentation in the iSAID dataset. First, the iSAID satellite images are split into 256 pixels by 256 pixels smaller images and then, the FreeSOLO model is applied to the smaller images. After the instance segmentation results are obtained, the predicted masks are merged together to form the mask of the shape of the original image.

The other two figures 4.6 and 4.7 show the visualization of the instance segmentation in two samples of AICrowd dataset. Here, inference is carried out in the images as they are. They are not split further because the images were medium resolution and they were not covering large area unlike the iSAID dataset. In this dataset too, the model is segmenting out other instances such as road, non-building structures, road curbs, and trees.

The three visualizations figure 4.11, figure 4.12, and figure 4.13 show the results of the FreeSOLO instance segmentation on the PASTIS dataset. Each image in this dataset has a dimension of 125 pixels by 125 pixels, which is comparatively smaller than other datasets. For this reason, the images are not divided further. The visualizations reveal the segmentation of the agricultural parcels, but it’s worth noting that a single segment may contain multiple parcels. The model appears to have difficulty distinguishing between separate parcels, unlike with everyday objects.
4.3.2 Quantitative Results

The table 4.1 shows the instance segmentation results in terms of average precision and average recall metrics. The FreeSOLO model is the class-agnostic unsupervised instance segmentation model so it segments instances of all the objects found in the image. However, the datasets have instances belonging to a limited number of classes. So, even if the FreeSOLO model is performing instance segmentation correctly on the category not contained in annotations, the evaluation flags that as a false positive. So, average precision scores are limited to single-digit figures. However, it is also to be noted that state-of-the-art class-agnostic unsupervised instance segmentation technique i.e. FreeSOLO on the COCO dataset has average precision scores in the single figure only. The relatively higher values of recall validates the hypothesis that precision is suffering because of incomplete instance annotations. Figure 4.8 shows the visualization of the instance segmentation on one of the iSAID images. As can be seen, the model is segmenting big trucks well but it is also segmenting a large portion of land which is not present in the annotations. The precision metric suffers some amount because of this too. The instances of PASTIS data did not have large objects (according to COCO definition) as the images themselves were of small dimension i.e., 128 pixel by 128 pixel. So, the average precision and average recall for large objects are given n/a value.

4.4 Transfer Learning

The next set of experiments was transfer learning on satellite images. I wanted to test the performance of image segmentation on labeled images, so the previously mentioned Dubai dataset was taken. I wanted to compare several su-
Table 4.1: Class-agnostic instance segmentation for iSAID and CrowdAI dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AP&lt;sub&gt;50:95&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;50&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;75&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;m50:95&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;50:95&lt;/sub&gt;</th>
<th>AR&lt;sub&gt;50:95&lt;/sub&gt;</th>
<th>AR&lt;sub&gt;m50:95&lt;/sub&gt;</th>
<th>AR&lt;sub&gt;l50:95&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>iSAID</td>
<td>0.4</td>
<td>0.9</td>
<td>0.5</td>
<td>0.9</td>
<td>1.2</td>
<td>2</td>
<td>6.3</td>
<td>12.8</td>
</tr>
<tr>
<td>CrowdAI</td>
<td>1.4</td>
<td>3.1</td>
<td>1.2</td>
<td>1.8</td>
<td>3.5</td>
<td>4.7</td>
<td>5.7</td>
<td>26</td>
</tr>
<tr>
<td>PASTIS</td>
<td>0.7</td>
<td>1.1</td>
<td>0.3</td>
<td>0.9</td>
<td>n/a</td>
<td>0.6</td>
<td>5</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Several sets of transfer learning experiments were conducted. At first, the pre-trained weights of several encoder-based backbones trained on ImageNet<sup>[9]</sup> database in a supervised learning fashion were taken<sup>[17][60][18][47][19][28]</sup>. They were used separately as feature extractors. Then, two semantic segmentation architectures i.e. feature pyramid network and U-Net<sup>[43]</sup> were trained in a supervised setting. The small labels Dubai dataset was used to train the model. The FreeSOLO backbone-based embeddings gave the comparative performance to the supervised learning based backbone weights on both the FPN and U-Net architectures. The results are tabulated in the table 4.2. The table has two metrics i.e. IOU score and dice coefficient tabulated for them on the test dataset. The segmentation results of the ResNet-101-based embeddings are shown in figure 4.14.

It is quite impressive that the model has learned to segment land, and water bodies. The data using which the weights were pre-trained is very much different than the satellite images. However, it is still able to do some semantic segmentation on the satellite images.
Table 4.2: SL vs SSL based encodings on semantic segmentation downstream task

<table>
<thead>
<tr>
<th>Encoder</th>
<th>Type</th>
<th>Decoder</th>
<th>Dice coefficient</th>
<th>IOU Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet18</td>
<td>Supervised</td>
<td>FPN</td>
<td>0.59</td>
<td>0.46</td>
</tr>
<tr>
<td>ResNet34</td>
<td>Supervised</td>
<td>FPN</td>
<td>0.57</td>
<td>0.44</td>
</tr>
<tr>
<td>ResNet50</td>
<td>Supervised</td>
<td>FPN</td>
<td>0.58</td>
<td>0.44</td>
</tr>
<tr>
<td>ResNet101</td>
<td>Supervised</td>
<td>FPN</td>
<td>0.6</td>
<td>0.48</td>
</tr>
<tr>
<td>ResNeXt50</td>
<td>Supervised</td>
<td>FPN</td>
<td>0.58</td>
<td>0.45</td>
</tr>
<tr>
<td>GERNet-S</td>
<td>Supervised</td>
<td>FPN</td>
<td>0.58</td>
<td>0.46</td>
</tr>
<tr>
<td>DenseNet121</td>
<td>Supervised</td>
<td>FPN</td>
<td>0.44</td>
<td>0.31</td>
</tr>
<tr>
<td>VGG11</td>
<td>Supervised</td>
<td>FPN</td>
<td>0.58</td>
<td>0.45</td>
</tr>
<tr>
<td>SE-Net</td>
<td>Supervised</td>
<td>FPN</td>
<td><strong>0.63</strong></td>
<td>0.5</td>
</tr>
<tr>
<td>FreeSOLO</td>
<td>Unsupervised</td>
<td>FPN</td>
<td>0.58</td>
<td>0.45</td>
</tr>
<tr>
<td>FreeSOLO</td>
<td>Unsupervised</td>
<td>U-Net</td>
<td><strong>0.62</strong></td>
<td><strong>0.50</strong></td>
</tr>
</tbody>
</table>
Similarly, the pre-trained weights of ResNet-101 architecture trained on ImageNet\cite{9} for FreeSOLO\cite{53} training are taken and used as a feature extractor. This feature extractor was trained in a self-supervised fashion, so, it did not require any labels. Then, the same set of architectures i.e. feature pyramid network and U-Net is used as above to carry out semantic segmentation on the satellite images. The comparable performance was achieved with respect to the IOU score and Dice coefficient on the held-out test dataset (shown in table 4.2.

It is impressive to see comparable performance of self-supervised learning based pre-trained weights compared to supervised learning pre-trained weights. This is advantageous because most of the time, the labels are not available, as obtaining labels is costly and time-consuming. So, a self-supervised learning framework can be used to learn the embeddings for satellite images. And, as shown above, comparable performance can be achieved in downstream tasks. The result produced by the SSL-based method is given in the figure 4.15.
Figure 4.2: Additional coarse masks extracted using free mask[53] on MBRSC dataset
Figure 4.3: Predicted masks using FreeSOLO[53] model on MBRSC dataset
Figure 4.4: Additional predicted masks using FreeSOLO[53] model on MBRSC dataset
Figure 4.5: FreeSOLO[53] predicted mask in PlanetScope UAH image

Figure 4.6: Instance segmentation in iSAID satellite image
Figure 4.7: Additional instance segmentation in iSAID satellite image

Figure 4.8: Class-agnostic instance segmentation in iSAID satellite image
Figure 4.9: Class-agnostic instance segmentation in a sample AICrowd satellite image

Figure 4.10: Additional class-agnostic instance segmentation in a sample AICrowd satellite image
Figure 4.11: Class-agnostic instance segmentation in a sample PASTIS satellite image

Figure 4.12: Additional class-agnostic instance segmentation in a sample PASTIS satellite image
Figure 4.13: Additional class-agnostic instance segmentation in a sample PASTIS satellite image

Figure 4.14: ImageNet\cite{9} supervised learning based pre-trained model on downstream segmentation task
Figure 4.15: ImageNet\cite{9} self-supervised learning based pre-trained model on downstream segmentation task
Chapter 5. Limitations and Applications

5.1 Limitations

The FreeSOLO model is a powerful method for instance segmentation, but it has several limitations that need to be addressed, especially when using it on other domains. Firstly, the DenseCL model used to generate free masks is not trained on domain-specific data, leading to domain gaps that also affect the FreeSOLO model. Additionally, as the model provides exhaustive instances of segmentation, it does not filter them to produce class-aware instance segmentation, making it less suitable for benchmarking in limited-class instance segmentation datasets. Furthermore, the SOLO architecture on which FreeSOLO is built suffers from the limitation of not being able to segment small instances though it does pretty well in segmenting the large regions of instances.

5.2 Applications

The FreeSOLO model, the state-of-the-art unsupervised class-agnostic instance segmentation method, can be followed by image classification to obtain panoptic segmentation. Panoptic segmentation has significant potential for use in satellite image analysis. By using instance segmentation to identify the instances and by using image classification to classify the objects in satellite images, FreeSOLO can produce metadata that can be used for a range of applications. For example, this approach can be used to generate metadata for satellite images that can be used to improve spatial search results, making it easier to locate specific
features or objects within an image, scene, or spatial boundary. Additionally, this method can be particularly useful for tracking changes over time, especially when using medium-resolution high-frequency data such as PlanetScope. By analyzing change detection over time, FreeSOLO combined with image classification can help researchers gain valuable insights into environmental and land-use changes.
Chapter 6. Conclusion and Future Research

6.1 Conclusion

This research work demonstrates the application of self-supervised learning based instance segmentation method to satellite images. The recent state-of-the-art methods for instance segmentation i.e. free mask and FreeSOLO[53] are taken and used to segment instances in satellite images. The methods are tested on two datasets: labeled Dubai[20] dataset and unlabeled PlanetScope[48] dataset of University of Alabama in Huntsville periphery. It is shown that the methods perform well in segmenting several objects in satellite images such as lakes, rivers, buildings, settlements, roads, and forests. For Dubai data[20], the segmentation results were validated by comparing them side-by-side with ground truth masks. As the Dubai data was labeled, the ground truth mask was available. For PlanetScope[48] unlabeled data, the segmentation results were validated visually as the university periphery was more familiar.

The benchmark for class-agnostic unsupervised instance segmentation was not found in the literature, so, the FreeSOLO model was benchmarked on three satellite images-based instance segmentation datasets, iSAID[56], CrowdAI[37], and PASTIS[44]. 0.9%AP$_{50}$ was achieved in the iSAID dataset, 3.1%AP$_{50}$ on the CrowdAI dataset, and 1.1%AP$_{50}$ on the PASTIS dataset. On large objects, the numbers increased to 1.2%AP$_{50}$ in the iSAID dataset and 3.5%AP$_{50}$ on the CrowdAI dataset. The state-of-the-art class-agnostic instance segmentation per-
formance on the COCO dataset is $AP_{50}$ of 9.8%. So, it is explainable given the limited number of categories in the annotations and domain gap.

The comparable performance of pre-trained weights trained using self-supervised learning i.e. FreeSOLO[53] was also shown with respect to the pre-trained weights trained using supervised learning on downstream semantic segmentation tasks on different semantic segmentation architecture. The Dubai[20] dataset was taken to quantitatively compare the pre-trained weights using Dice coefficient and IOU score metric. The self-supervised learning based pre-trained weights gave IOU score of 0.50 and Dice coefficient of 0.62 compared to supervised learning-based pre-trained weights (best IOU score of 0.48 and Dice coefficient of 0.63). This is advantageous because labels were not available for massively available PlanetScope[48] satellite data. The embeddings can be pre-trained in a self-supervised learning fashion as described in DenseCL[53] and FreeSOLO[53], and later used in downstream tasks.

6.2 Future Research

Currently, the coarse masks are extracted through the DenseCL[54] model. The DenseCL[54] model is pre-trained on ImageNet[9] database. The database contains the images of everyday objects. However, these methods are being applied and tested on satellite images. So, in the future, similar model such as DenseCL[54] can be pre-trained in just satellite images. This is also possible because of unsupervised nature of it. It does not need any labels to do self-supervised learning. So, the first line of research work can be learning a DenseCL[54] based pre-trained model on just satellite images. While extracting the free masks, currently the score maps are extracted by finding cosine similarity of all the queries with all the keys. This requires the free mask approach to do filtering of the
masks using NMS method. This methodology might possibly be improved if SWIN transformer[32] architecture can be used because SWIN transformer makes use of shifted-window attention. This methodology naturally might provide non-redundant masks as the attention mechanism is applied on each window.

The second line of research can be training a weakly supervised learning-based model on top of free masks extracted by the satellite images pre-trained model. This can be a similar work to that of FreeSOLO[53] but specific to satellite images with satellite images focused architecture. Better segmentation models could be obtained by this method. This is certainly interesting and impactful research that could be done in the future.

In this research, transfer learning experiments are done by extracting features from the pre-trained ResNet-101 backbone. All the weights of the decoder part have been learned for doing semantic image segmentation. In the future, SOLOv2[55] architecture can be used to perform semantic image segmentation and also reused for the decoder network too.
References


[48] Planet Team. Planet application program interface: In space for life on earth, 2017–.


