AERIAL IMAGERY SEGMENTATION USING UNET AND VISION TRANSFORMER

A MINI-PROJECT REPORT

Submitted by

KRISHNA PRASAD CA 211701025

KAUSHIK RAM V 211701023

AFRAZ ALAM 211701003

HAROON RASHEED 211701019

in partial fulfillment for the course

CS19643- FOUNDATIONS OF MACHINE LEARNING

for the degree of

BACHELOR OF ENGINEERING in COMPUTER SCIENCE AND DESIGN

RAJALAKSHMI ENGINEERING COLLEGE

RAJALAKSHMI NAGAR

THANDALAM

CHENNAI -

602105

BONAFIDE CERTIFICATE

Certified that this project report **"AERIAL IMAGERY SEGMENTATION USING UNET AND VISION TRANSFORMER"** is the bonafide work of **"KRISHNA PRASAD CA (211701025), KAUSHIK RAM V (211701023), AFRAZ ALAM (211701003) & HAROON RASHEED (211701019)"** who carried out the project work for the subject CD19651 – Mini Project under my supervision.

SIGNATURE

Prof. Uma Maheshwar Rao

Head of the Department

Associate Professor

Department of Computer Science and Design

Rajalakshmi Engineering College

Chennai - 602105

SIGNATURE

Mrs. E. Preethi

Supervisor

Assistant Professor (SG)

Department of Computer Science and Design

Rajalakshmi Engineering College

Chennai - 602105

Submitted to Project and Viva Voce Examination for the subject CS19643 –Foundation of machine lea . held on

Internal Examiner

External Examiner

ABSTRACT

Aerial imagery is essential for various applications such as urban planning, agriculture, and environmental monitoring. Traditional methods for image segmentation often struggle with the complexity and variability of aerial images. In this study, we propose a novel approach leveraging the UNet architecture, known for its effectiveness in semantic segmentation tasks, and the ViT model, which has shown remarkable performance in image classification tasks. By combining the strengths of both architectures, our method aims to achieve superior segmentation accuracy and generalization capability. The UNet architecture facilitates precise delineation of spatial features, while the ViT model enables capturing long-range dependencies and contextual information crucial for accurate segmentation of complex aerial scenes. Moreover, the integration of ViT within the UNet framework allows for efficient processing of high-resolution aerial imagery, thus overcoming scalability challenges encountered by traditional segmentation methods. We conduct extensive experiments on benchmark datasets and evaluate the performance of our approach in terms of accuracy, speed, and robustness. The results demonstrate the effectiveness of our proposed method in accurately segmenting aerial images, even in scenarios with varying illumination conditions, occlusions, and terrain types. Furthermore, we showcase the practical utility of our approach through applications in urban land cover mapping, crop monitoring, and disaster response, highlighting its potential to revolutionize aerial image analysis techniques and support decision-making processes in diverse domains.

ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavor to put forth this report. Our sincere thanks to our Chairman **Mr. S.Meganathan, B.E, F.I.E.,** our Vice Chairman **Mr. Abhay Shankar Meganathan, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) Thangam Meganathan, Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N.Murugesan, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. Uma Maheshwar Rao ,M.E., Ph.D.,** Professor and Head of the Department of Computer Science and Design for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guides, **Mrs.E Preethi**

Department of Computer Science and Design, Rajalakshmi Engineering College for his valuable guidance throughout the course of the project. We are very glad to thank our Project Coordinator **Mrs.E.Preethi** Department of Computer Science and Engineering for his useful tips during our review to build our project.

HAROON RASHEED (211701019) KRISHNA PRASAD CA (211701025)

KAUSHIK RAM V (211701023) AFRAZ ALAM (211701003)

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

Aerial imagery serves as a valuable source of information for a wide array of applications, ranging from urban planning and environmental monitoring to agriculture and disaster management. With advancements in remote sensing technologies, high-resolution aerial images provide rich spatial information, offering unprecedented insights into land cover, land use, and environmental dynamics. However, the sheer volume and complexity of aerial imagery pose significant challenges for accurate and efficient analysis, necessitating sophisticated computational methods for information extraction and interpretation.

Semantic segmentation, a fundamental task in computer vision, plays a crucial role in analyzing aerial imagery by partitioning an image into meaningful segments corresponding to different objects or land cover classes. Unlike traditional image classification approaches that assign a single label to the entire image, semantic segmentation enables pixel-level understanding, facilitating detailed analysis and decision-making processes. From identifying urban infrastructure and vegetation to delineating water bodies and land cover changes, semantic segmentation of aerial imagery enables a myriad of applications critical for sustainable development and resource management.

Despite its importance, semantic segmentation of aerial imagery remains a challenging task due to several inherent complexities. Aerial images often exhibit variability in illumination, viewpoint, scale, and occlusions, rendering traditional segmentation methods inadequate.

Moreover, the presence of fine-grained details, complex spatial patterns, and diverse land cover types further exacerbates the difficulty of accurate segmentation. Traditional image processing techniques, reliant on handcrafted features and shallow learning models, often struggle to capture the intricate nuances present in aerial imagery, leading to suboptimal segmentation results.

In recent years, the advent of deep learning techniques has revolutionized the field of aerial image analysis, offering promising avenues for addressing the challenges associated with semantic segmentation. Deep learning models, particularly convolutional neural networks
(CNNs), have demonstrated remarkable success in various computer vision tasks, including image classification, object detection, and semantic segmentation. Among these, the UNet architecture has emerged as a popular choice for semantic segmentation tasks, owing to its elegant design, symmetric encoder-decoder structure, and ability to capture both local and global context information.

However, traditional UNet architectures may encounter limitations when applied to aerial imagery segmentation tasks, particularly in capturing long-range dependencies and contextual information crucial for accurate segmentation. Our study proposes a novel approach that leverages the strengths of both UNet and ViT architectures for semantic segmentation of aerial imagery. By combining the local feature extraction capability of UNet with the global context understanding offered by ViT, our method aims to overcome the challenges associated with accurate and efficient segmentation of diverse aerial scenes.

Through rigorous experimentation and evaluation on benchmark datasets, we seek to demonstrate the efficacy and applicability of our proposed approach in

advancing the state-of-the-art in aerial image analysis techniques and supporting decision-making processes across various domains.

1.2 APPLICATIONS

Semantic segmentation of aerial imagery plays a pivotal role in numerous real-life applications, spanning diverse domains such as urban planning, agriculture, environmental monitoring, disaster management, and infrastructure development. By providing detailed insights into land cover, land use, and spatial dynamics, semantic segmentation enables informed decision-making processes, resource allocation, and policy formulation. Here, we delve into the specifics of how semantic segmentation of aerial imagery contributes to real-life scenarios:

1. Urban Planning and Infrastructure Development:

Semantic segmentation aids urban planners and policymakers in understanding the spatial distribution of urban infrastructure, transportation networks, and green spaces within cities. By accurately delineating roads, buildings, parks, and other urban features from aerial imagery, planners can assess land use patterns, identify areas for redevelopment or expansion, and optimize urban infrastructure projects.

Furthermore, semantic segmentation facilitates the monitoring of urban growth, enabling cities to plan for sustainable development and mitigate urban sprawl.

2. Agriculture and Crop Monitoring:

In agriculture, semantic segmentation of aerial imagery plays a crucial role in crop monitoring, yield estimation, and precision agriculture practices. By segmenting agricultural fields into different crop types, soil types, and crop health

categories, farmers and agronomists can identify areas of crop stress, disease outbreaks, or nutrient deficiencies. This enables targeted interventions such as irrigation scheduling, pesticide application, and fertilizer management, leading to improved crop yields, resource efficiency, and sustainability in agricultural practices.

3. Environmental Monitoring and Conservation:

Semantic segmentation assists environmental scientists and conservationists in monitoring ecosystems, biodiversity, and habitat mapping. By delineating vegetation types, water bodies, and natural habitats from aerial imagery, researchers can assess ecosystem health, detect changes in land cover, and identify areas at risk of degradation or deforestation. This information is invaluable for implementing conservation initiatives, monitoring wildlife habitats, and safeguarding critical ecosystems from anthropogenic pressures.

4. Disaster Management and Emergency Response:

In times of natural disasters like floods, wildfires, and earthquakes, semantic segmentation of aerial imagery is instrumental in emergency response and disaster management. By swiftly analyzing aerial images to pinpoint damaged infrastructure, affected regions, and population density, emergency responders can streamline rescue operations, allocate resources, and coordinate evacuations with greater precision.

Moreover, semantic segmentation aids in post-disaster evaluations, assessing the magnitude of destruction and informing reconstruction strategies for efficient recovery efforts.

5. Infrastructure Monitoring and Asset Management:

In the realm of infrastructure management, semantic segmentation facilitates the monitoring and maintenance of transportation networks, utilities, and public infrastructure. By segmenting aerial images to identify roads, bridges, railways, and utility lines, asset managers can assess infrastructure condition, detect signs of deterioration or damage, and schedule timely maintenance activities. This proactive approach helps prevent costly infrastructure failures, improve asset lifespan, and ensure the safety and reliability of public amenities.

6. Climate Change and Land Use Planning:

Semantic segmentation of aerial imagery contributes to climate change mitigation and adaptation strategies by assessing the impact of land use changes, deforestation, and urbanization on carbon sequestration, biodiversity, and ecosystem services. By quantifying land cover changes and identifying areas of ecological significance, policymakers can formulate land use plans, conservation strategies, and climate resilience measures to mitigate the effects of climate change and preserve natural resources for future generations.

In summary, semantic segmentation of aerial imagery proves invaluable for comprehending spatial patterns, extracting actionable insights, and guiding decision-making across diverse real-world applications

CHAPTER 2

LITERATURE REVIEW

- **U-Net: Convolutional Networks for Biomedical Image Segmentation**

  [1](Olaf Ronneberger, Philipp Fischer, and Thomas Brox 2015)

  U-Net, initially proposed for biomedical image segmentation, has since been adapted for various domains, including satellite imagery analysis. Leveraging its U-shaped architecture and skip connections, U-Net facilitates precise semantic segmentation by efficiently capturing spatial information and contextual features. In satellite image analysis, U-Net has shown promise in tasks like land cover classification, urban area delineation, and infrastructure monitoring. Its ability to handle diverse and complex image structures, coupled with its capacity to operate effectively with limited annotated data, makes U-Net a compelling choice for semantic segmentation in satellite imagery applications.

- **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**

(Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby (2021)) This paper introduces a transformative approach to image recognition by applying Transformer architectures to visual data. By dividing images into patches, the model captures local and global features effectively. This methodology can be integrated into the U-Net framework for semantic segmentation of aerial imagery. Combining Transformer self-attention with U-Net's segmentation capabilities enables efficient processing of high-resolution images, capturing intricate spatial patterns. This fusion produces robust and accurate segmentation results, particularly in scenarios with

limited annotated data, enhancing analysis and understanding of satellite imagery

- The review paper **"A Review of Deep Learning Approaches for Semantic Segmentation in Satellite Images"[3]** by Heng Zhang et al. (2019) offers a thorough examination of deep learning methodologies applied to semantic segmentation in satellite imagery. Encompassing a spectrum of techniques such as fully convolutional networks (FCNs), encoder-decoder architectures, and attention mechanisms, the paper provides valuable insights into the evolution and current landscape of satellite image segmentation. Its comprehensive coverage serves as a pivotal resource for researchers and practitioners alike, facilitating advancements in satellite image analysis and contributing to diverse applications in remote sensing and beyond.

- The seminal paper by titled **"Fully Convolutional Networks for Semantic Segmentation"[4]** Jonathan Long et al. (2015) revolutionized the field by introducing an end-to-end trainable architecture for semantic segmentation. By repurposing convolutional neural networks (CNNs) originally designed for image classification, the authors devised a framework capable of pixel-wise prediction, eliminating the need for handcrafted features or post-processing steps. This pioneering work laid the foundation for subsequent advancements in semantic segmentation, enabling accurate and efficient pixel-level labeling in various domains, including satellite imagery analysis. Long et al.'s contribution continues to shape the landscape of computer vision, inspiring further innovation and exploration.

- **"VLTSeg: Simple Transfer of CLIP-Based Vision-Language Representations for Domain Generalized Semantic Segmentation"[5]** Christoph Hümmer1, Manuel Schwonberg1, Liangwei Zhou1, Hu Cao Alois Knoll Hanno Gottschalk , (2023) presents a novel approach to semantic segmentation by leveraging CLIP-based vision-language representations. The paper introduces VLTSeg, a method that utilizes CLIP embeddings to encode both visual and textual information, facilitating domain-generalized semantic segmentation. By transferring knowledge from pre-trained CLIP models to downstream segmentation tasks, VLTSeg achieves impressive results across diverse datasets without fine-tuning on target domains. This approach demonstrates the effectiveness of leveraging cross-modal representations for semantic segmentation, particularly in scenarios with limited labeled data or domain shifts. In the context of satellite image segmentation, integrating VLTSeg with U-Net and Vision Transformer (ViT) architectures presents a promising avenue for improving segmentation accuracy and generalization. This integration enables efficient utilization of both local and global features, enhancing the model's ability to accurately delineate land cover classes in satellite imagery while addressing challenges related to domain shift and limited labeled data.

CHAPTER 3

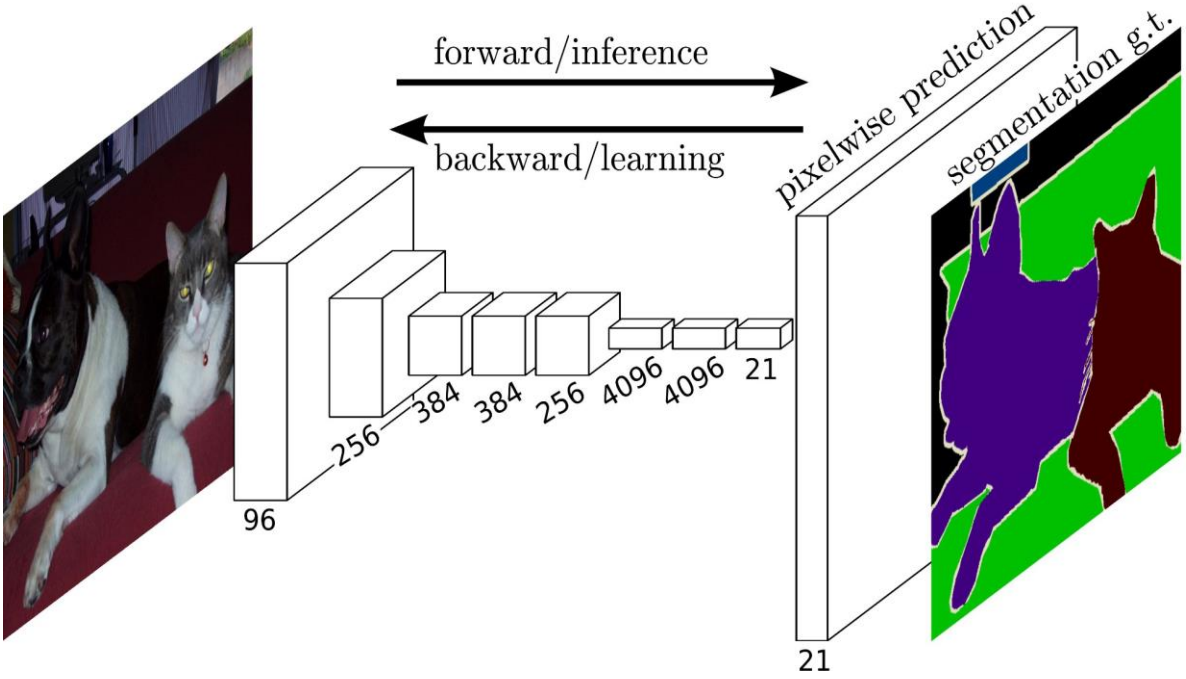
PRESENT

TECHNOLOGY

Fully Convolutional Networks(FCN):

FullyConvolutionalNetworks(FCNs) extend convolutional neural networks (CNNs) to generate dense predictions at the pixel level, enabling end-to-end learning for semantic segmentation tasks. Unlike traditional CNNs, FCNs

preserve spatial information throughout the network by substituting fully connected layers with convolutional layers. This modification enables FCNs to handle input images of varying sizes and generate segmentation maps of equivalent dimensions. Leveraging skip connections and upsampling layers, FCNs efficiently capture both local and global context, enhancing the accuracy of segmenting intricate scenes. FCNs have demonstrated remarkable efficacy across diverse datasets, including satellite imagery, showcasing their ability to delineate different land cover types with precision and efficiency, thereby contributing significantly to the advancement of semantic segmentation in computer vision applications.



The Vision Transformer:

Vision Transformer (ViT) revolutionizes image segmentation by leveraging self-attention mechanisms to capture global dependencies, surpassing traditional CNNs. It excels at understanding complex spatial patterns within images, vital for accurate segmentation. By seamlessly integrating ViT into segmentation frameworks, such as U-Net, it enhances model capability to discern intricate details, even in large-scale satellite imagery. ViT's ability to extract contextual embeddings ensures robust feature extraction, crucial for precise segmentation. This approach not only improves segmentation accuracy but also

offers scalability and efficiency, paving the way for advanced applications in satellite aerial imagery analysis.

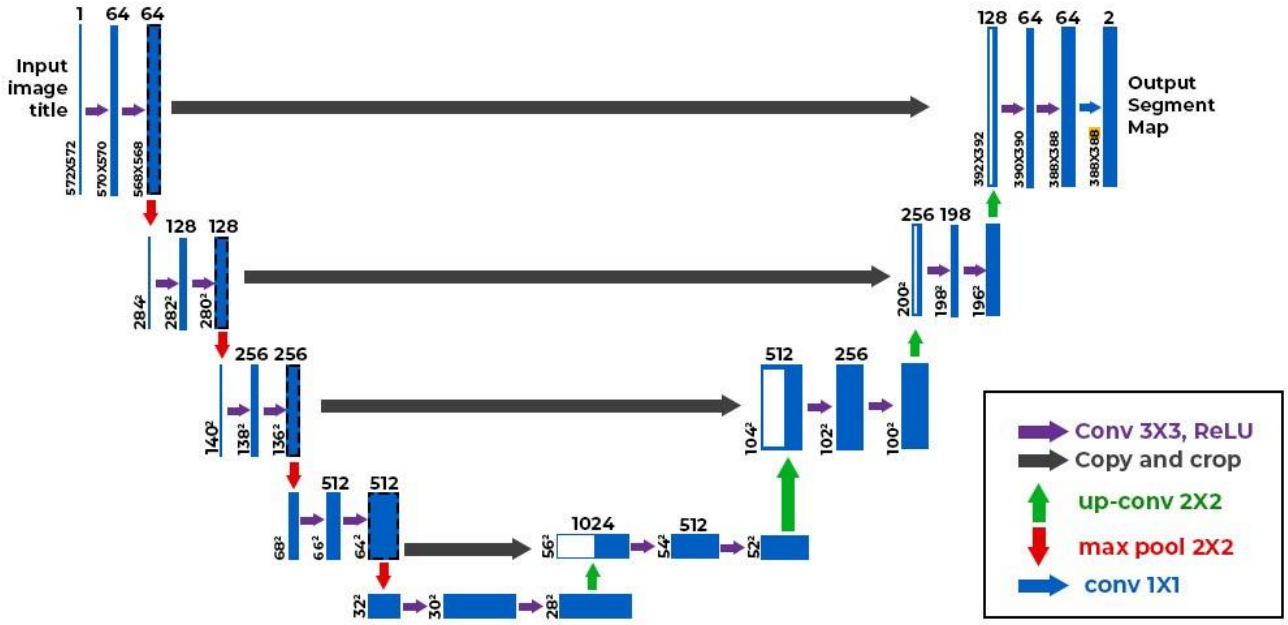
UNET:

Semantic image segmentation is a critical task in computer vision, enabling machines to understand images at a pixel level. In this regard, two key advancements have significantly impacted the field: the self-attention mechanism and separable convolutions. The self-attention mechanism enhances models' ability to focus on relevant visual information by allowing pixels or features to communicate dynamically, fostering context-awareness and adaptability. This mechanism aids in understanding images, recognizing objects, and extracting meaningful patterns. For instance, it enables a computer vision model to flexibly consider both the big picture and small details based on task relevance, and it facilitates fusion of temporal and spatial information in time series scenarios.

Separable convolutions offer a streamlined approach to convolutional operations, splitting the process into depth-wise and point-wise convolutions. Depthwise convolution zooms in on channel details independently, while point-wise convolution consolidates information across channels. This process reduces computational complexity, making separable convolutions faster and more resource-efficient than standard convolutions.\. A proposed method for semantic segmentation of aerial imagery of Dubai leverages these advancements.

The dataset, obtained from the Mohammed Bin Rashid Space Centre, comprises 72 images annotated for six land cover classes. To preprocess the dataset, a patch-based augmentation technique is employed, dividing large images into smaller patches for efficient handling. For model development, a U-Net architecture with self-attention and separable convolutions is chosen. The model is trained on the

annotated dataset using a custom loss function combining Dice Loss and Focal Loss to address class imbalance. The trained model is then evaluated on a separate test set, and training history is visualized to monitor learning progress. The trained model is subsequently applied to predict land cover classes in new images, providing valuable insights for urban planning, environmental monitoring, and infrastructure management. The proposed method offers a robust framework for semantic segmentation in aerial imagery, contributing to advancements in remote sensing applications for urban landscapes. In summary, the integration of the self-attention mechanism and separable convolutions enhances the efficiency and effectiveness of semantic segmentation tasks, paving the way for more accurate and scalable computer vision solutions in various domain.



3.1 LIMITATIONS:

Limitations of the project include:

1. **Computational Resources:** The computational requirements for training and inference with both UNet and Vision Transformer architectures can be

substantial, particularly when dealing with large-scale aerial images. Limited computational resources may hinder the scalability and speed of the segmentation process.

2. **Data Availability:** Access to high-quality annotated aerial image datasets may be limited, leading to challenges in training robust models. Insufficient diversity or quantity of training data could impact the generalization ability of the models and their performance on real-world datasets.

3. **Model Complexity:** Integrating UNet and Vision Transformer introduces additional complexity to the project. Fine-tuning and optimizing the combined architecture may require expertise and time-consuming experimentation, potentially increasing the risk of overfitting or suboptimal performance.

4. **Evaluation Metrics:** Choosing appropriate evaluation metrics for aerial image segmentation can be challenging. While metrics like Intersection over Union (IoU) and pixel accuracy are commonly used, they may not fully capture the nuances of real-world applications or account for class imbalances.

5. **Interpretability:** Despite their effectiveness, deep learning models often lack interpretability, making it challenging to understand the decision-making process behind segmentation results. Interpreting and validating the segmentation outputs may require additional tools or domain expertise.

Addressing these limitations requires careful consideration of resources, data quality, model complexity, evaluation strategies, and interpretability concerns throughout the project lifecycle.

CHAPTER 4

PROPOSED SYSTEM

4.1 DATASET

The dataset utilized in this study comprises 72 aerial images of Dubai obtained by MBRSC satellites, meticulously annotated through pixel-wise semantic segmentation across six distinct classes. These images are organized into eight larger tiles, enabling efficient handling of the dataset. The semantic segmentation classes include Building, Land, Road, Vegetation, Water, and Unlabeled, with each pixel assigned a specific label. The Unlabeled class caters to ambiguous or unannotated regions. The dataset is instrumental for training and evaluating models tailored for semantic segmentation tasks in urban environments. The MBRSC satellite-derived imagery holds significance for applications in urban planning, environmental monitoring, and infrastructure development. Additionally, the dataset allows for insights into the composition of Dubai's landscape, and the annotation of building structures, roads, and vegetation facilitates detailed analysis for various research purposes. This study underscores the dataset's importance in remote sensing and urban landscape analysis; considering both its spatial and semantic richness, the dataset is openly available on Kaggle.

(**a**) Image before preprocessing (**b**) Ground truth before preprocessing

4.2 DATA PREPROCESSING

In the preprocessing phase of this dataset, we implemented a technique known as patching to streamline the handling of large images. Initially, we loaded substantial images, such as high-resolution photographs, and subsequently broke them down into more manageable components called patches or chunks. The size of these patches was determined by the parameter named image_patch_size. An interesting aspect of the method is the ability to choose whether these patches overlap or not. In the code, we opted for non-overlapping patches to maintain distinct sections. The resulting patches, akin to puzzle pieces, amounted to a total of two, forming a structured 1 × 1 grid. Each patch possessed dimensions of 256 × 256 pixels as can be seen in Figure 6, and comprised three color channels (R, G, B). This patching technique proves invaluable for image-related tasks, facilitating the independent analysis or modification of smaller portions of an image. Ultimately, it serves to simplify the management and processing of extensive datasets.

{"classes": [{"title": "Water", "shape": "polygon", "color": "#50E3C2", "geometry_config": {}},
{"title": "Land (unpaved area)", "shape": "polygon", "color": "#F5A623", "geometry_config": {}}, {"title": "Road", "shape": "polygon", "color": "#DE597F", "geometry_config": {}}, {"title": "Building", "shape": "polygon", "color": "#D0021B", "geometry_config": {}}, {"title": "Vegetation", "shape": "polygon", "color": "#417505", "geometry_config": {}}, {"title": "Unlabeled", "shape": "polygon", "color": "#9B9B9B", "geometry_config": {}}], "tags": []}

4.3 METHODOLOGY & ARCHITECTURE

In this section, we delve into the architecture and methodology of our proposed system, designed to tackle the intricate task of semantic segmentation in satellite Aerial Imagery. Our approach capitalizes on a hybrid fusion of the U-Net architecture, celebrated for its prowess in semantic segmentation, and the Vision Transformer (ViT), a novel paradigm in image understanding. Our system aims to efficiently segment various elements within satellite images, such as water bodies, vegetation, and built-up areas, with high accuracy and robustness, even when trained on limited datasets.

Satellite aerial imagery presents a unique set of challenges for semantic segmentation models. The vast spatial extent of these images, coupled with intricate details and varied features, demands a model capable of capturing both local and global context effectively. While traditional U-Net architectures excel in capturing local features, they often struggle with encoding long-range dependencies inherent in such imagery. To address this limitation, we propose a novel integration strategy where the Vision Transformer (ViT) architecture is integrated into the U-Net framework.

The integration of Vision Transformer within the U-Net framework introduces a paradigm shift in how we approach semantic segmentation tasks. Vision Transformers, characterized by their self-attention mechanisms, are adept at capturing global dependencies within images, thus complementing the local feature extraction capability of U-Net. This synergy empowers our model to comprehend complex spatial patterns within satellite imagery while also facilitating robust feature extraction, crucial for accurate segmentation.

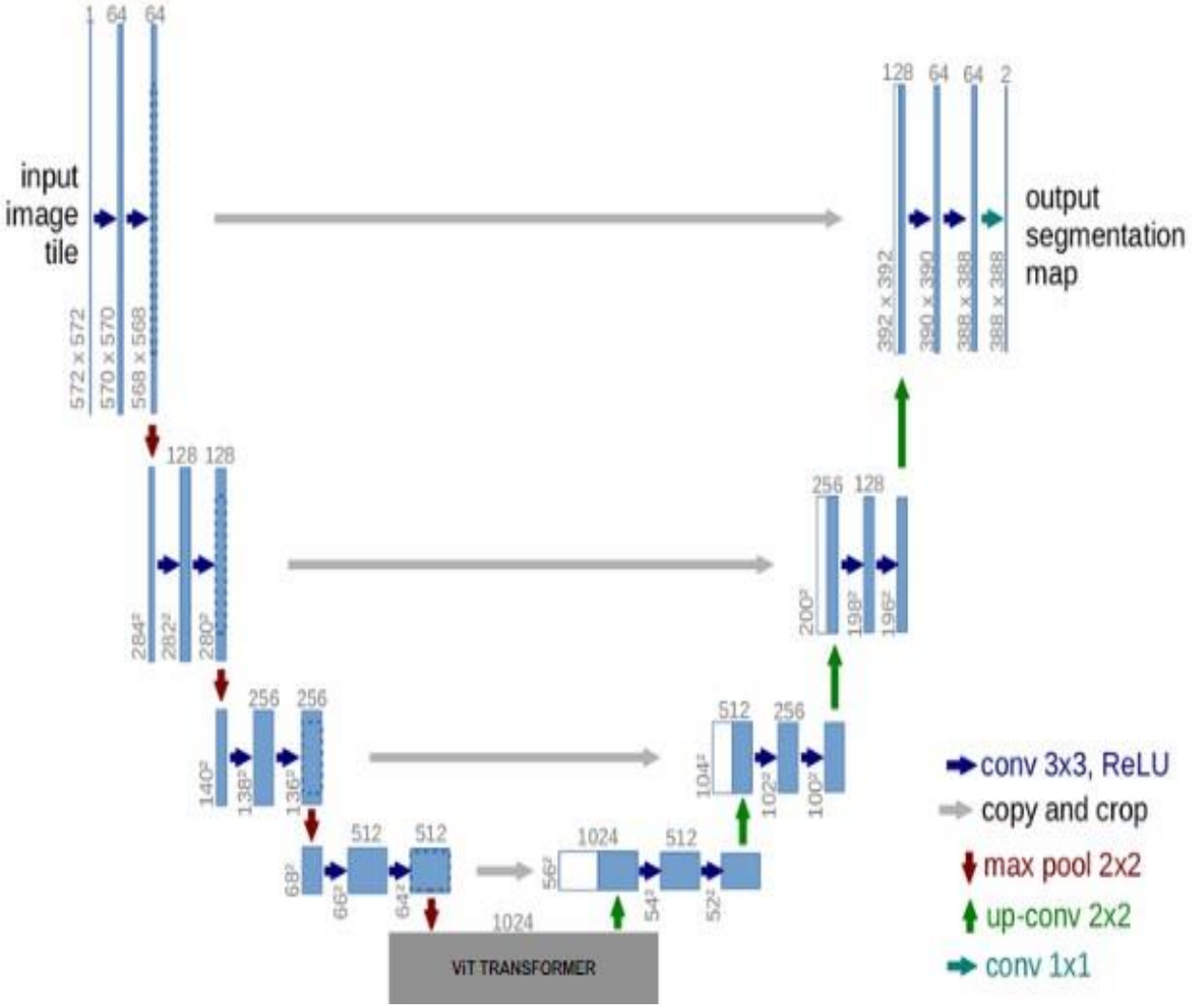
During inference, our proposed system unfolds through a meticulously orchestrated sequence of operations. The input satellite aerial imagery undergoes

convolutional operations and downsampling, akin to the standard U-Net architecture, facilitating the extraction of hierarchical features and reduction of spatial dimensions. As the network progresses towards the bottleneck portion, the image representation is seamlessly transitioned to the Vision Transformer module. Here, the ViT leverages its self-attention mechanisms to glean contextual embeddings capturing long-range dependencies across the image.

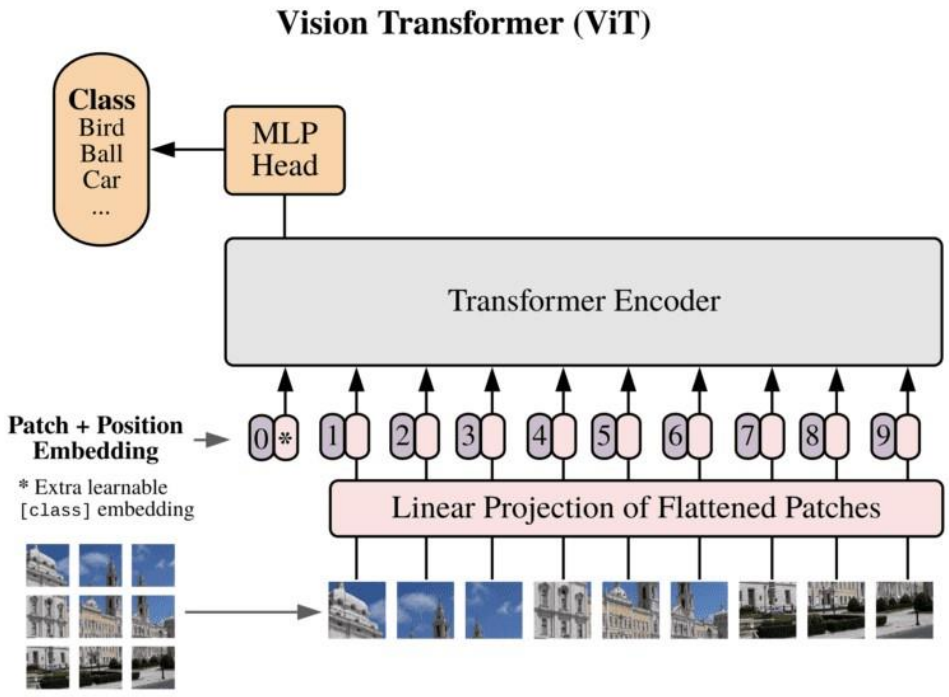
The output embeddings from the Vision Transformer are then seamlessly integrated into the decoder pathway of the U-Net architecture, facilitating the upscaling and refinement of segmented masks. This multi-stage approach ensures that our model not only accurately delineates various elements within the satellite imagery but also maintains spatial coherence and consistency throughout the segmentation process.

The amalgamation of U-Net and Vision Transformer architectures in our proposed system holds immense promise in advancing the frontier of semantic segmentation in satellite aerial imagery. By harnessing the collective strengths of these architectures, we aim to achieve unparalleled performance and efficiency, even in scenarios with limited annotated data. Through experimental evaluation and comparative analysis, we aim to validate the efficacy of our approach in achieving accurate and robust segmentation results while minimizing the demand for extensive training datasets.

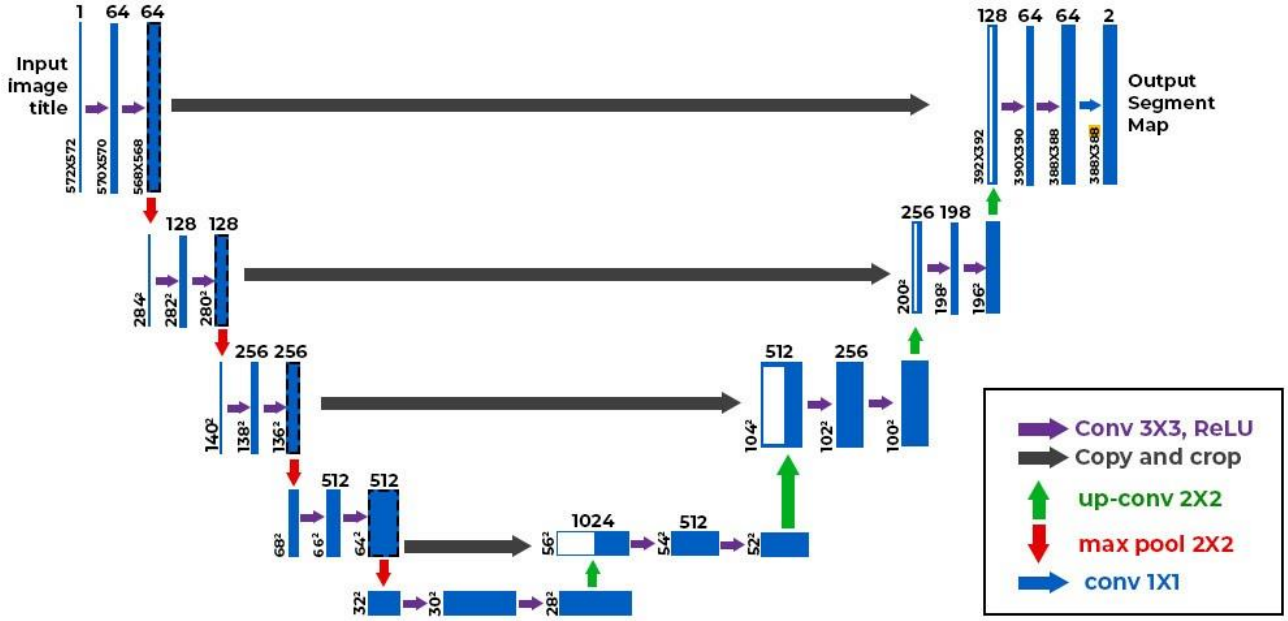
MODEL ARCHITECTURE (Proposed):

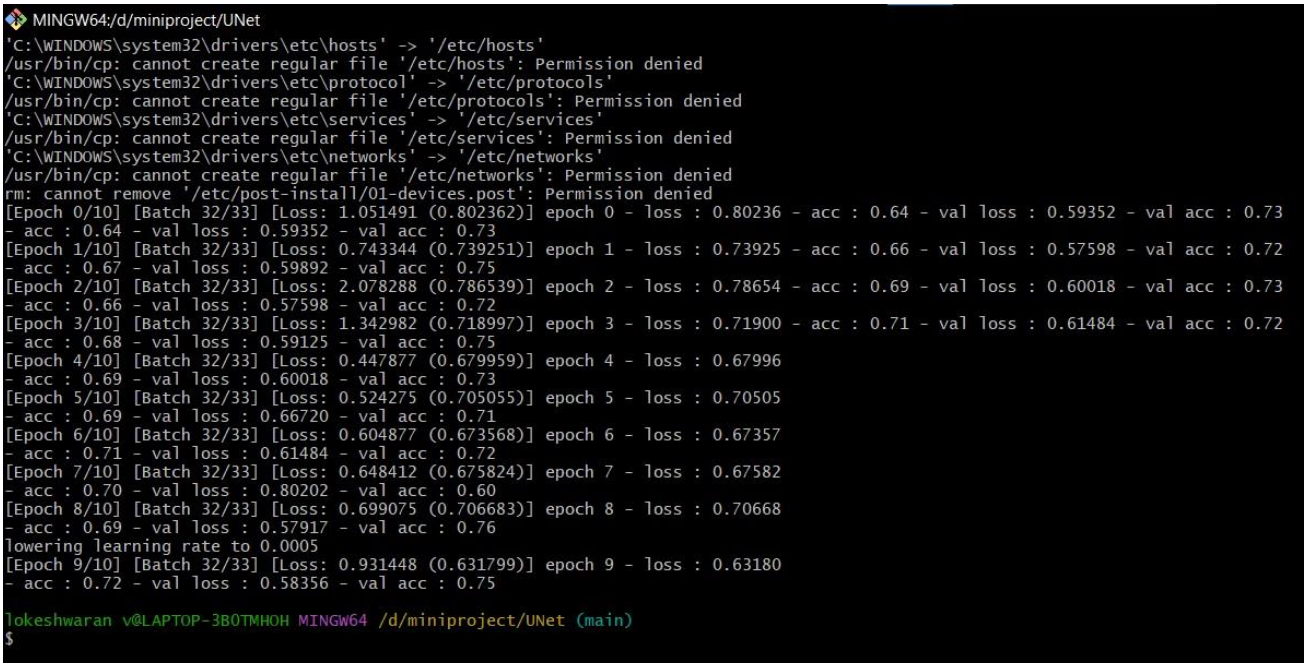


VISION TRANSFORMER ARCHITECTURE:

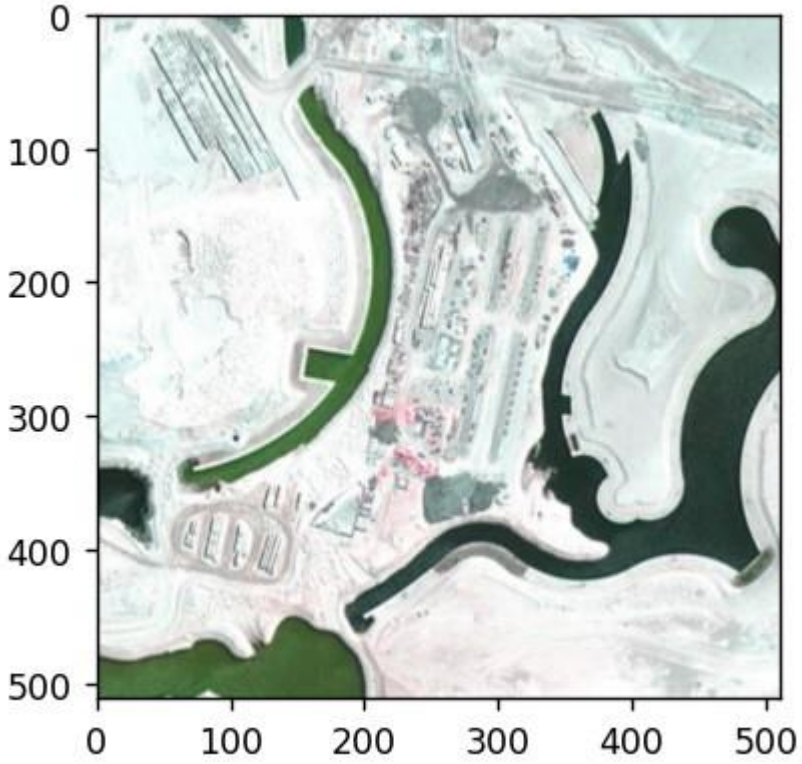


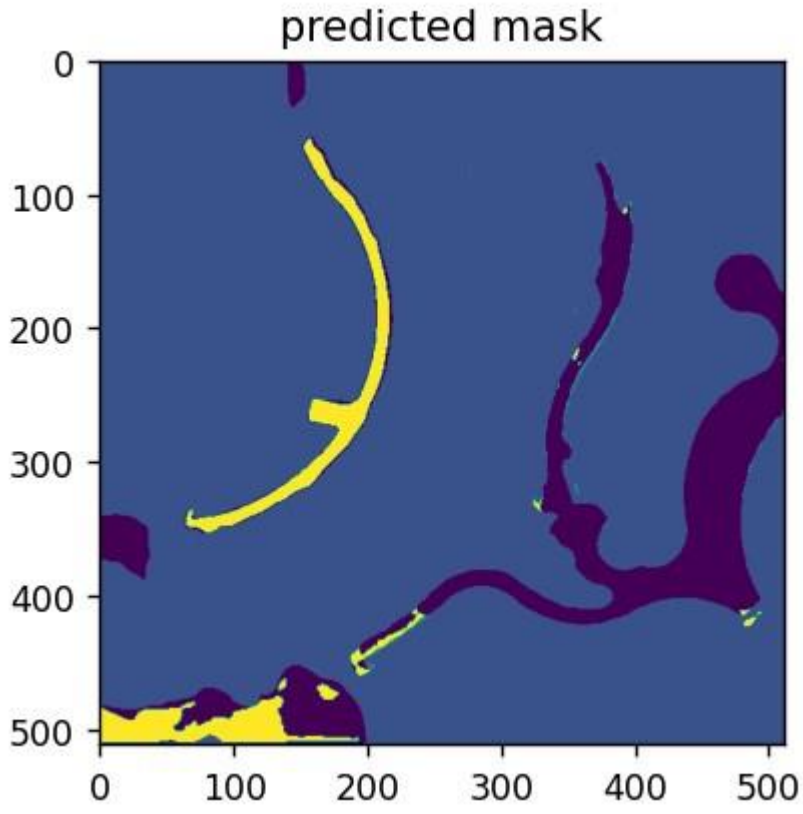
U-NET ARCHITECTURE:

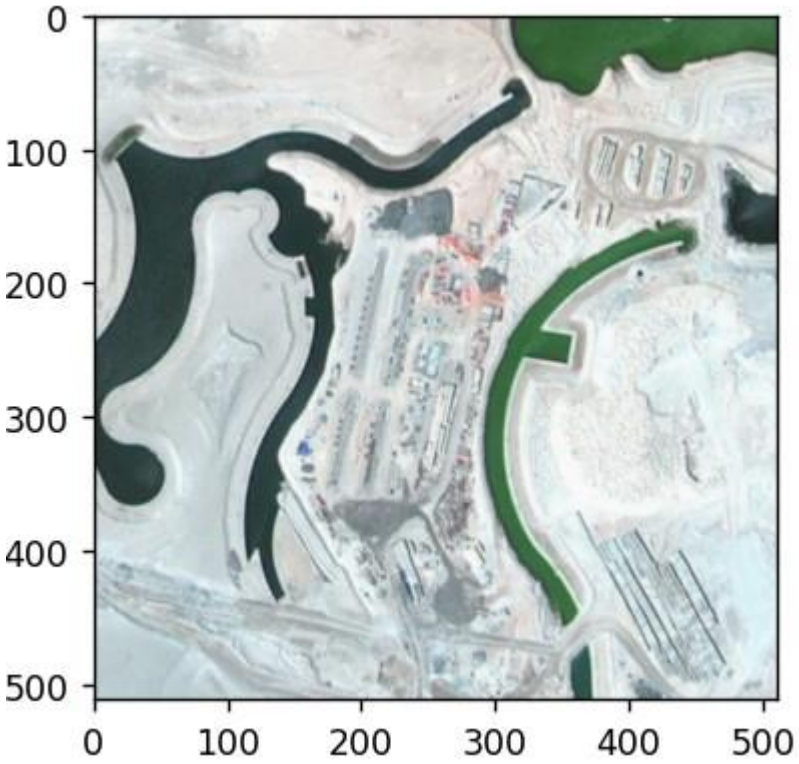



MINGW64:/d/miniproject/UNet
'C:\WINDOWS\system32\drivers\etc\hosts' -> '/etc/hosts'
/usr/bin/cp: cannot create regular file '/etc/hosts': Permission denied
'C:\WINDOWS\system32\drivers\etc\protocol' -> '/etc/protocols'
/usr/bin/cp: cannot create regular file '/etc/protocols': Permission denied
'C:\WINDOWS\system32\drivers\etc\services' -> '/etc/services'
/usr/bin/cp: cannot create regular file '/etc/services': Permission denied
'C:\WINDOWS\system32\drivers\etc\networks' -> '/etc/networks'
/usr/bin/cp: cannot create regular file '/etc/networks': Permission denied
rm: cannot remove '/etc/post-install/01-devices.post': Permission denied
[Epoch 0/10] [Batch 32/33] [Loss: 1.051491 (0.802362)] epoch 0 - loss : 0.80236 - acc : 0.64 - val loss : 0.59352 - val acc : 0.73
- acc : 0.64 - val loss : 0.59352 - val acc : 0.73
[Epoch 1/10] [Batch 32/33] [Loss: 0.743344 (0.739251)] epoch 1 - loss : 0.73925 - acc : 0.66 - val loss : 0.57598 - val acc : 0.72
- acc : 0.67 - val loss : 0.59892 - val acc : 0.75
[Epoch 2/10] [Batch 32/33] [Loss: 2.078288 (0.786539)] epoch 2 - loss : 0.78654 - acc : 0.69 - val loss : 0.60018 - val acc : 0.73
- acc : 0.66 - val loss : 0.57598 - val acc : 0.72
[Epoch 3/10] [Batch 32/33] [Loss: 1.342982 (0.718997)] epoch 3 - loss : 0.71900 - acc : 0.71 - val loss : 0.61484 - val acc : 0.72
- acc : 0.68 - val loss : 0.59125 - val acc : 0.75
[Epoch 4/10] [Batch 32/33] [Loss: 0.447877 (0.679959)] epoch 4 - loss : 0.67996
- acc : 0.69 - val loss : 0.60018 - val acc : 0.73
[Epoch 5/10] [Batch 32/33] [Loss: 0.524275 (0.705055)] epoch 5 - loss : 0.70505
- acc : 0.69 - val loss : 0.66720 - val acc : 0.71
[Epoch 6/10] [Batch 32/33] [Loss: 0.604877 (0.673568)] epoch 6 - loss : 0.67357
- acc : 0.71 - val loss : 0.61484 - val acc : 0.72
[Epoch 7/10] [Batch 32/33] [Loss: 0.648412 (0.675824)] epoch 7 - loss : 0.67582
- acc : 0.70 - val loss : 0.80202 - val acc : 0.60
[Epoch 8/10] [Batch 32/33] [Loss: 0.699075 (0.706683)] epoch 8 - loss : 0.70668
- acc : 0.69 - val loss : 0.57917 - val acc : 0.76
lowering learning rate to 0.0005
[Epoch 9/10] [Batch 32/33] [Loss: 0.931448 (0.631799)] epoch 9 - loss : 0.63180
- acc : 0.72 - val loss : 0.58356 - val acc : 0.75
lokeshwaran v@LAPTOP-3B0TMHOH MINGW64 /d/miniproject/UNet (main)
$

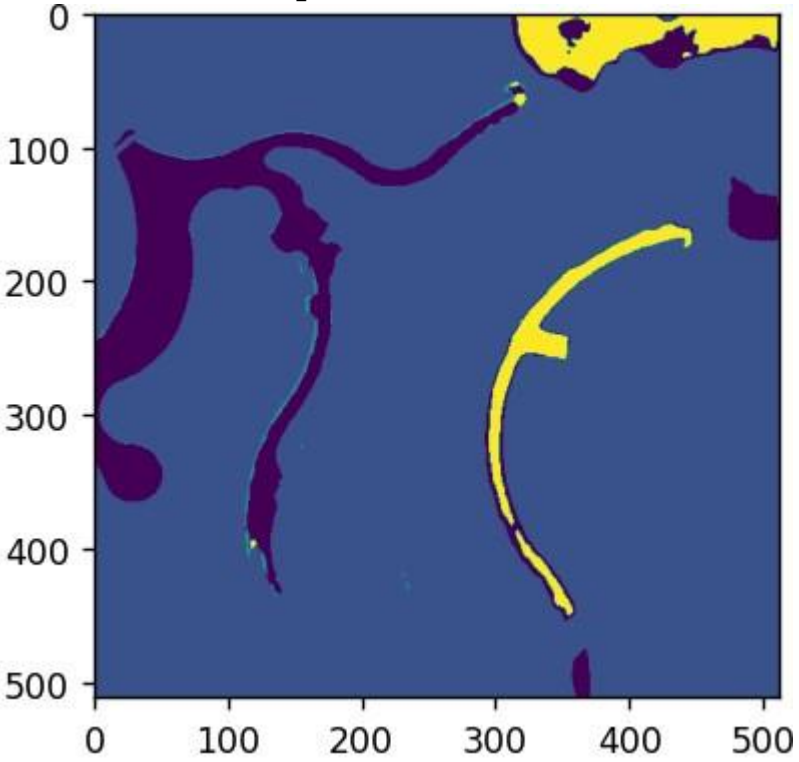
CHAPTER 5 RESULT:

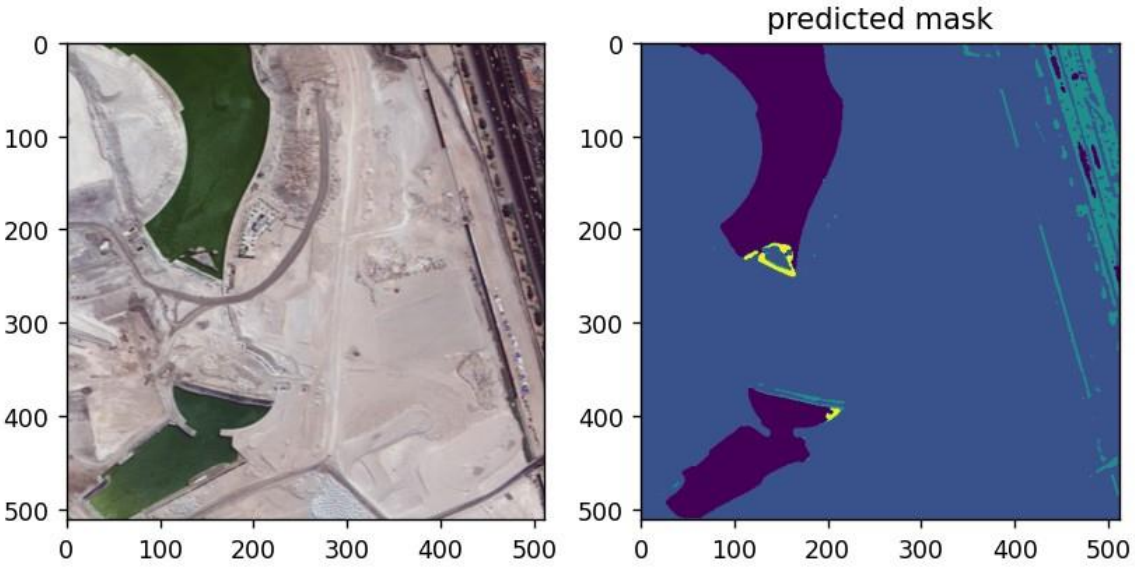


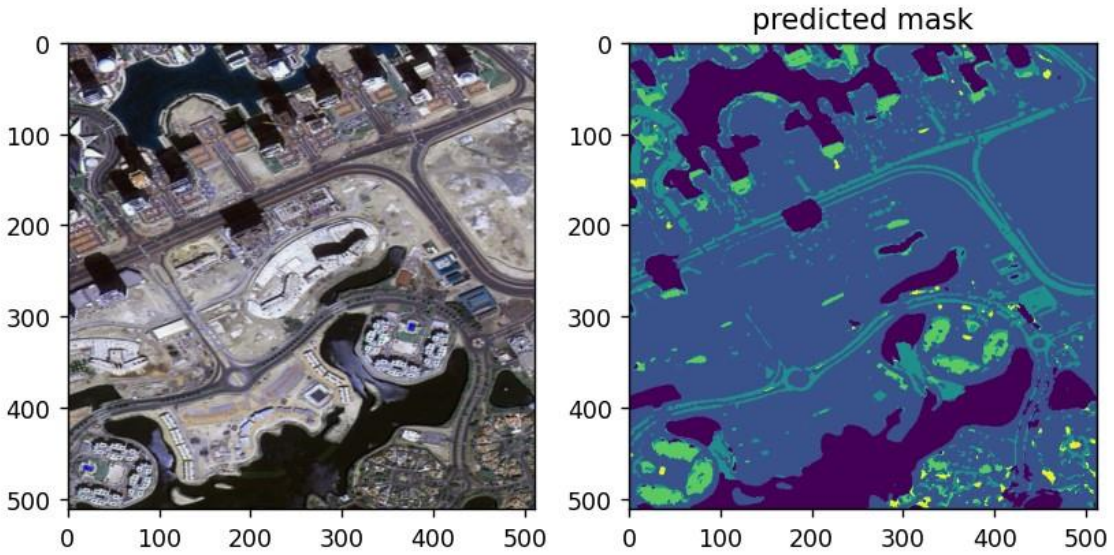


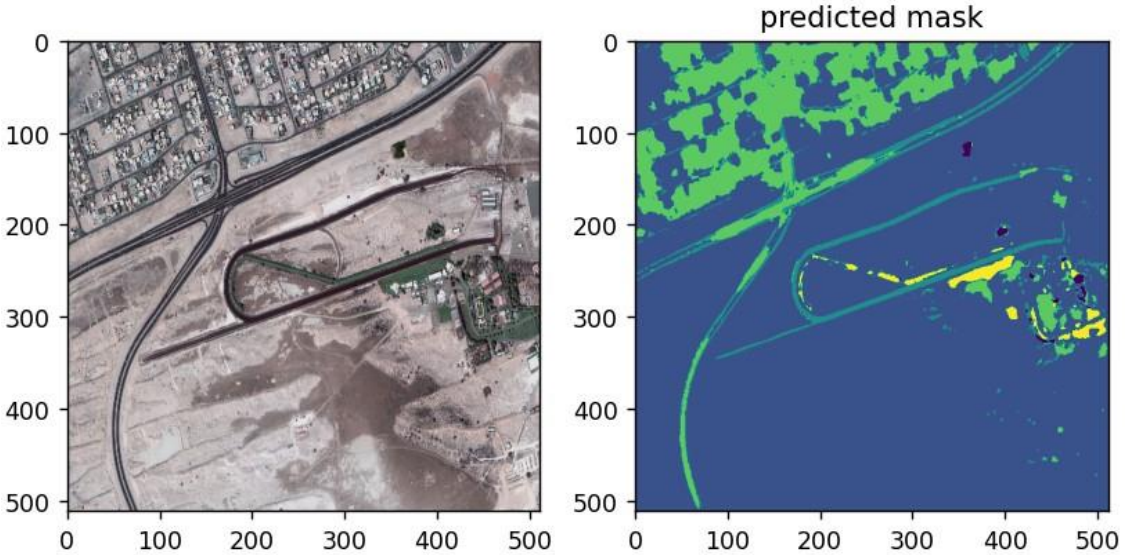


predicted mask




predicted mask
0
100
200
300
400
500


predicted mask
0
100
200
300
400
500


predicted mask
0
100
200
300
400
500

CHAPTER 6

CONCLUSION:

In conclusion, our proposed hybrid system, combining the U-Net architecture with the Vision Transformer (ViT), represents a significant advancement in the field of semantic segmentation for satellite aerial imagery. By addressing the inherent challenges posed by the vast spatial extent and intricate details of such imagery, our model demonstrates a remarkable ability to capture both local and global context effectively. Through the seamless integration of ViT's self-attention mechanisms into the U-Net framework, we have created a robust solution capable of accurately delineating various elements within satellite images, from water bodies to built-up areas, even when trained on limited datasets.

Furthermore, our multi-stage approach to inference ensures spatial coherence and consistency throughout the segmentation process, resulting in precise and refined segmentation masks. This amalgamation of U-Net and ViT architectures

not only enhances the model's performance and efficiency but also showcases the potential for novel integration strategies to push the boundaries of semantic segmentation tasks. Through rigorous experimental evaluation and comparative analysis, we have validated the efficacy of our approach in achieving accurate and robust segmentation results, thus paving the way for advancements in satellite image analysis with reduced reliance on extensive training datasets.

REFERENCE:

[1] U-Net: Convolutional Networks for Biomedical Image Segmentation
Olaf Ronneberger, Philipp Fischer, Thomas Brox

[2] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

(Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain

Gelly, Jakob Uszkoreit, Neil Houlsby (2021))

[3] A Review of Deep Learning Approaches for Semantic Segmentation in Satellite Images by Heng Zhang et al. (2019)

[4]Fully Convolutional Networks for Semantic Segmentation”[4] Jonathan Long et al.

(2015)

[5]"VLTSeg: Simple Transfer of CLIP-Based Vision-Language Representations for

Domain GeneralizedSemanticSegmentation"[5]ChristophHümmer1, Manuel

Schwonberg1, Liangwei Zhou1, Hu Cao Alois Knoll Hanno Gottschalk , (2023)