Multimodal Dual-Swin Transformer with Spectral-Spatial Feature Extraction for Terrain Recognition

Team Spectra Transformers

1 Introduction

While traditional Deep Learning methods like CNN can effectively handle Multiclass Terrain classification, the estimation of implicit properties demands a more sophisticated approach, making it a primary focus. In Remote sensing, terrain roughness is traditionally assessed using Lidar (Light Detection and Ranging) to create Digital Elevation Maps (DEM).

However, Lidar has limitations, including environmental distortion and a relatively short range, rendering it unsuitable for satellite-based defence applications. Hyperspectral Imaging offers a solution to this problem by providing rich spatial and spectral resolution, making it resilient to environmental distortion.

Hyperspectral Image Classification is a task in the field of remote sensing and computer vision. It involves the classification of pixels in hyperspectral images into different classes based on their spectral signature.

Hyperspectral images contain information about the reflectance of objects in hundreds of narrow, contiguous wavelength bands, making them useful for a wide range of applications, including mineral mapping, vegetation analysis, and urban land-use mapping.

Primary objective of this project is to accurately identify and classify different terrains present given image. This involves sandy, rocky, grass, marshy, etc. terrains and their classification based on their spatial & spectral properties.

This project proposal provides a detailed system methodology for implementing the terrain recognition system, along with implicit properties estimation.

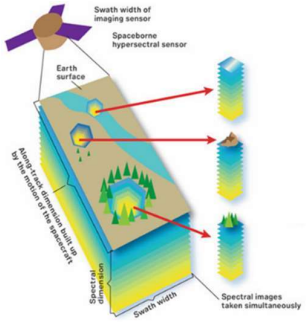

Figure 1: Hyperspectral Data Acquisition

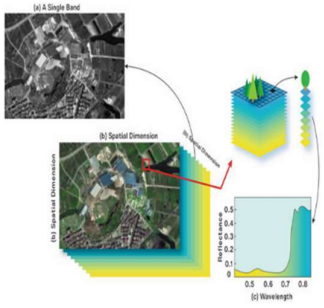

Figure 2: Hyperspectral Data Cube

2 Problem Statement

Deep Learning for Terrain Recognition - Vision based methods using deep learning such as CNN to perform terrain recognition (sandy/rocky/grass/marshy) enhanced with implicit quantities information such as the roughness, slipperiness, an important aspect for high-level environment perception.

3 Scope

The proposed system will be able to accurately classify terrain types like sandy, rocky, marshy, and grassy along with estimating implicit properties such as roughness and slipperiness.

For RGB input, the desired output includes the predicted terrain class, coupled with estimated roughness and slipperiness achieved through spatial analysis.

In the case of HSI input, the desired output is the predicted terrain class and implicit properties information such as roughness, slipperiness, and the spectral signature for each pixel or patch using spatial and spectral analysis.

Implicit Properties should be visualized using a colormap overlay over the original image.

4 Proposed Solution

We intend to develop a robust terrain classification system using RGB and Hyperspectral Images. The primary objective is to leverage advanced feature extraction and deep learning techniques to accurately classify different terrains and estimate implicit properties such as roughness and slipperiness, which can be achieved by using advanced computer vision and remote sensing techniques.

In our prototype, we have used SWIN (shifted window) image transformer for terrain recognition, and statistical texture analysis for roughness and slipperiness estimation, using RGB dataset.

Our goal is to develop an application, which can take either RGB or Hyperspectral multimodal image as input, and give image terrain class and pixel-level implicit properties information as output.

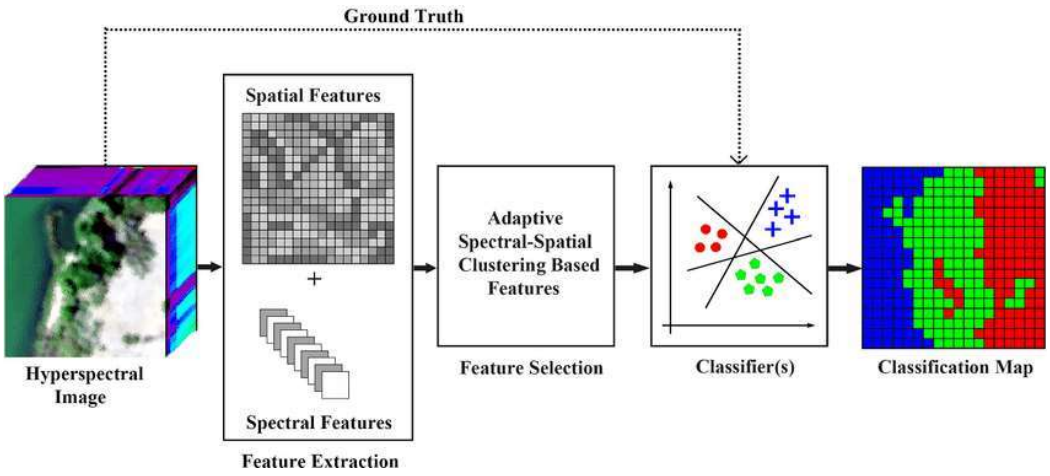

Figure 3: Proposed System Block Diagram

4.1 Dataset

In this project, we intend to acquire and prepare a diverse dataset of 3D hyperspectral data cubes for terrain recognition. These data cubes will serve as our primary data source, capturing both spectral and spatial information across numerous contiguous bands. We aim to include various terrains, encompassing sandy, rocky, grassy, and marshy areas, under an array of lighting and weather conditions.

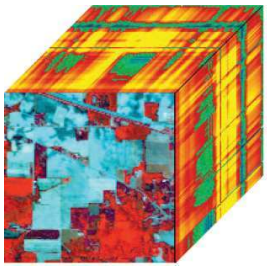

Figure 4: HSI 3D Data Cube

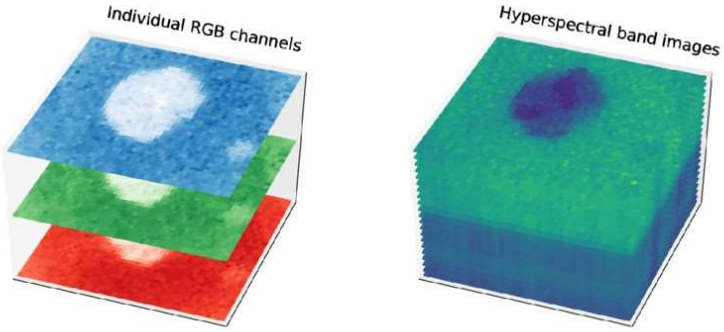

Figure 5: RGB vs HSI Images

4.2 Feature Extraction

Our methodology will revolve around spatial approach for RGB images and a spectral-spatial hybrid feature extraction approach for HSI images. This approach combines spectral information derived from reflectance values across bands with spatial information obtained from pixel arrangements within the hyperspectral data cubes.

By doing so, we aim to enhance the model's capability to classify terrains accurately, and reduce its computational complexity.

4.3 Feature Selection

In our project, we propose the utilization of an Adaptive Spectral-Spatial Clustering-based feature selection algorithm. This sophisticated algorithm is designed to meticulously identify and select the most pertinent features from the hyperspectral data cubes, ensuring that the model receives the most salient and discriminative information for accurate terrain recognition.

By harnessing the power of spectral and spatial clustering techniques, this feature selection process will contribute significantly to enhancing the model's ability to discern subtle spectral signatures and spatial patterns within the data. Through this approach, we aim to optimize the feature space, improve classification performance, and ultimately advance the precision of our terrain recognition system.

4.4 Classification

The cornerstone of our project lies in the design and implementation of Dual-Branch Swin Transformer, based on RGB and HSI input.

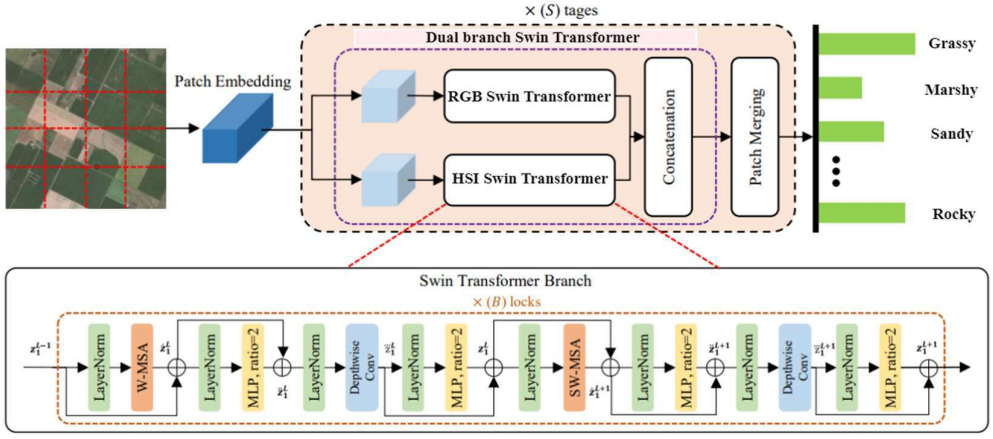

Figure 6: Model Architecture

4.5 Output

For RGB input, the desired output includes the predicted terrain class, coupled with estimated roughness and slipperiness achieved through spatial analysis. In the case of HSI input, the desired output is the predicted terrain class and implicit properties information such as roughness, slipperiness, and the spectral signature for each pixel or patch using spatial and spectral analysis.

5 Project Prototype

The project prototype, which is trained on an RGB Dataset, achieves a test accuracy of 99% for the terrain recognition task, and terrain roughness and slipperiness information is calculated using a statistical variance-based patch texture analysis and is visualized as an overlay over the original image.

The model was trained on a dataset consisting of about 45.1k images, with more than 10k images for each terrain class (Grassy, Marshy, Rocky, Sandy).

The prototype can be viewed here: https://github.com/MaitreyaShelare/Spectra-Transformers-SIH-2023

Following was the training history

```
Epoch 0/9
----------
100%|██████████| 247/247 [02:31<00:00,  1.63it/s]
train Loss: 0.4909 Acc: 0.9455
100%|██████████| 212/212 [00:30<00:00,  6.97it/s]
val Loss: 0.4088 Acc: 0.9854

Epoch 1/9
----------
100%|██████████| 247/247 [02:10<00:00,  1.89it/s]
train Loss: 0.4096 Acc: 0.9871
100%|██████████| 212/212 [00:26<00:00,  7.90it/s]
val Loss: 0.3912 Acc: 0.9923

Epoch 2/9
----------
100%|██████████| 247/247 [02:11<00:00,  1.89it/s]
train Loss: 0.3962 Acc: 0.9921
100%|██████████| 212/212 [00:26<00:00,  7.97it/s]
val Loss: 0.3815 Acc: 0.9941

Epoch 3/9
----------
100%|██████████| 247/247 [02:10<00:00,  1.90it/s]
train Loss: 0.3904 Acc: 0.9941
100%|██████████| 212/212 [00:26<00:00,  7.94it/s]
val Loss: 0.3800 Acc: 0.9950

Epoch 4/9
----------
100%|██████████| 247/247 [02:10<00:00,  1.89it/s]
train Loss: 0.3855 Acc: 0.9960
100%|██████████| 212/212 [00:26<00:00,  7.99it/s]
val Loss: 0.3778 Acc: 0.9953

Epoch 5/9
----------
100%|██████████| 247/247 [02:11<00:00,  1.88it/s]
train Loss: 0.3835 Acc: 0.9960
100%|██████████| 212/212 [00:26<00:00,  8.02it/s]
```

```
val Loss: 0.3746 Acc: 0.9965

Epoch 6/9
----------
100%|██████████| 247/247 [02:10<00:00,  1.89it/s]
train Loss: 0.3813 Acc: 0.9969
100%|██████████| 212/212 [00:26<00:00,  8.00it/s]
val Loss: 0.3756 Acc: 0.9963

Epoch 7/9
----------
100%|██████████| 247/247 [02:10<00:00,  1.89it/s]
train Loss: 0.3809 Acc: 0.9970
100%|██████████| 212/212 [00:26<00:00,  7.97it/s]
val Loss: 0.3742 Acc: 0.9967

Epoch 8/9
----------
100%|██████████| 247/247 [02:10<00:00,  1.89it/s]
train Loss: 0.3790 Acc: 0.9978
100%|██████████| 212/212 [00:26<00:00,  8.03it/s]
val Loss: 0.3711 Acc: 0.9976

Epoch 9/9
----------
100%|██████████| 247/247 [02:10<00:00,  1.89it/s]
train Loss: 0.3769 Acc: 0.9982
100%|██████████| 212/212 [00:26<00:00,  7.99it/s]
val Loss: 0.3713 Acc: 0.9972

Training complete in 26m 38s
Best Val Acc: 0.9976
```

Following was the test accuracy

```
Test Loss: 0.0009
Test Accuracy of Grassy: 99% (1814/1818)
Test Accuracy of Marshy: 99% (1641/1650)
Test Accuracy of Rocky: 99% (1640/1643)
Test Accuracy of Sandy: 100% (1641/1641)
Test Accuracy of 99% (6736/6752)
```

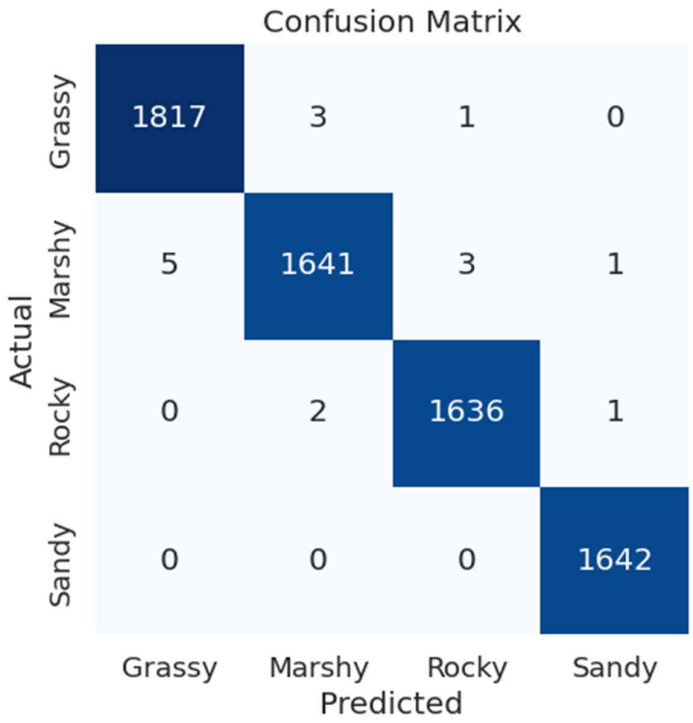

Figure 7: Confusion Matrix

Following is the implicit properties estimation using statistical variance-based patch texture analysis

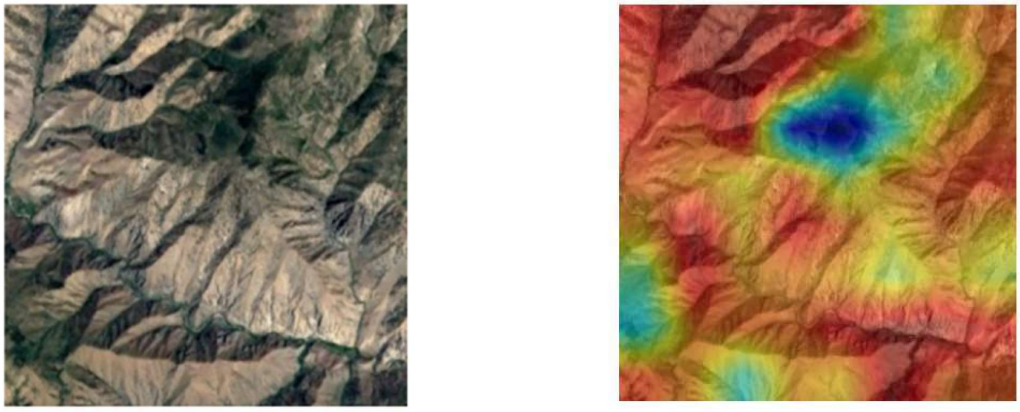

Figure 8: Implicit Properties estimation