



Department of Electrical & Computer Engineering

North South University

CSE 468.1

Spatio-Temporal Adaptive Fusion Transformer (STAFT) for next

Group Members:

Name	NSU ID
Khan Mohammad Sashoto Seeam	1831769642
Nesar Ahmed	2211836642
Ahnaf Mohammed Mahi Kabir	2222171042

Supervisor:

Dr. Md Adnan Arefeen [AFE]

Assistant Professor, Department of Electrical & Computer Engineering

Introduction

Land Cover Classification (LCC) is the process of identifying different surface types on Earth, such as water, vegetation, and urban areas, using satellite images. Although deep learning models like U-Net and DeepLab have shown strong results in image segmentation, they often struggle with the complex patterns and multispectral features found in satellite images. Traditional CNNs use fixed filters that cannot easily adapt to variations in texture, scale, and spectral information (like RGB or NIR), which leads to inaccurate classification in large or diverse areas. Transformer-based models like SegFormer can capture global context but are computationally demanding.

This project addresses the challenge of **improving pixel-wise land-cover segmentation accuracy** while maintaining **computational efficiency**. The core idea is to enhance U-Net's feature-extraction capability through **Selective Kernel ResNeXt (SK-ResNeXt)** modules, which adaptively adjust receptive fields and combine multi-scale feature responses.

Related Work

This project aims to reproduce and extend the model proposed in “*Leveraging U-Net and Selective Feature Extraction for Land Cover Classification Using Remote Sensing Imagery*,” implementing a U-Net with an SK-ResNeXt encoder to enhance feature extraction and improve segmentation accuracy on multispectral datasets.

Methods

Our first goal is to reproduce the first paper and then experiment with both traditional and transformer-based approaches. The original paper achieved a +6 % Overall Accuracy (OA) and a +8 % mIoU gain over the vanilla U-Net. It didn't handle class imbalance. The author has concluded that further improvement is possible by improving the decoder. Moreover, incorporating architectures that were not included can enhance the result.

Implementation Details (Baseline Model – U-Net + SK-ResNeXt-50)

Backbone: SK-ResNeXt-50

Loss function: Cross-Entropy Loss (optionally Dice Loss for class imbalance)

Optimizer: Adam (learning rate = 5×10^{-4})

Input size: 256×256 pixel patches

Training strategy: 80% training, 20% validation split

Data augmentation: Random flips & rotations

Variant A – SK-ResNeXt Encoder + Attention Gate Decoder

Backbone: SK-ResNeXt-50 (ImageNet-pretrained)

Decoder: U-Net decoder with Attention Gates inserted before each skip connection

Loss function: Weighted Cross-Entropy + Dice Loss (weights derived from class frequency).

Optimizer: AdamW

Input size: 256×256 px patches (4 channels: R, G, B, NIR).

Data augmentation: Random flips, rotations, brightness, and contrast jitter.

Variant B – SK-ResNeXt Encoder + Attention Gate Decoder

Everything is the same as Variant A, except for the following.

Backbone: SK-ResNeXt-50 (ImageNet-pretrained)

Decoder: Dense (U-Net++)-style decoder where each upsampling stage receives inputs from all previous encoder and decoder levels (via dense skip paths).

Loss function: Deep Supervision with Multi-output Dice + Cross-Entropy Loss

Variant C – Hybrid (CNN + ViT Fusion)

Parallel SK-ResNeXt-50 and ViT-Tiny or SwinT encoders with an attention-based fusion layer.

Proposed Framework

We would like to go ahead with **PyTorch** as there are plenty of research work available on public domain with PyTorch.

Evaluation

The evaluation of the proposed U-Net with SK-ResNeXt encoder for Land Cover Classification (LCC) relies on a large-scale dataset and established semantic segmentation metrics.

Dataset

The method is evaluated on the Five-Billion-Pixels dataset. This is a large-scale land cover dataset featuring 150 high-resolution satellite images (Gaofen-2) with over 5 billion manually annotated pixels. The imagery consists of four spectral bands: Blue, Green, Red, and Near-Infrared (NIR), enabling multispectral analysis across 24 distinct land cover categories.

Evaluation Metrics

Performance is quantified using three standard metrics:

Overall Accuracy (OA): The simplest measure, indicating the proportion of all correctly classified pixels in the dataset.

Intersection over Union (IoU): Measures the precision of the segmentation boundary by calculating the overlap between the predicted area and the actual ground-truth area for a single class.

Mean Intersection over Union (mIoU): The average IoU across all 24 land cover classes. This is the most crucial metric, as it provides a balanced assessment of the model's performance across all categories, including those that are rare or complex.

Training/Inference time: Since one of our goals is to reduce complexity and avoid using full ViT, we will also evaluate training time.