Technical Report: Weakly Supervised Remote Sensing Segmentation

1. Introduction & Problem Statement

Image segmentation traditionally classifies each pixel in an image, requiring complete, dense segmentation masks during training. However, in real-world remote sensing scenarios, annotation is often based on points or incomplete tagging, representing a significant deep learning challenge. To address this limited point annotation information, we developed a custom deep learning framework designed to learn effectively from extremely sparse data.

2. Method

To solve the point-labeling constraint, the pipeline was constructed using the following approaches:

- **Baseline Architecture:** We utilized transfer learning via a pre-trained model (U-Net with a ResNet-34 backbone initialized on ImageNet weights) to guarantee a strong feature-extraction baseline.
- **Simulated Point Labels:** We sourced a remote sensing image segmentation dataset and randomly sampled the dense masks to simulate sparse point labels, keeping only 100 marked pixels per image. All unmarked pixels were assigned an "ignore" index. This semi-supervised approach aims to make full use of the data to improve model performance.
- **Partial Cross-Entropy (pfCE) Loss:** A special loss function was implemented to allow point labeling to train the segmentation model. The $pfCE$ function computes the pixel-wise Focal Loss, masks out all unknown pixels, and calculates the average error exclusively over the known points using the formula:

  pfCE = ΣMASK_labeled (FocalLoss(pred, GT) × MASK_labeled) / ΣMASK_labeled

3. Experiment: Training Convergence Under Sparse Annotations

Purpose & Hypothesis The objective of this experiment was to explore the training stability and performance of a deep learning model when penalized only by a sparse subset of pixels (100 points per image patch). The hypothesis states that the custom pfCE loss will successfully provide enough gradient signal for the model to converge and accurately segment dense, irregular geographical features over time.

Experimental Process

1. **Data:** Used the "Semantic segmentation of aerial imagery" dataset containing aerial tiles and 6 classes (Water, Land, Road, Building, Vegetation, Unlabeled).
2. **Simulation:** Dense ground truth masks were aggressively reduced to 100 random points per 256x256 tensor.

3. **Training:** The model was trained using the Adam optimizer (Learning Rate = 0.001) for 5 epochs. The $pfCE$ loss was added to the network to monitor performance.

Results

The experiment successfully validated the hypothesis. Despite utilizing only a fraction of the available annotation data, the model exhibited highly stable convergence. Over 5 epochs, the Average Loss decreased consistently:

- **Epoch 1:** 0.2450
- **Epoch 2:** 0.1540
- **Epoch 3:** 0.1379
- **Epoch 4:** 0.1313
- **Epoch 5:** 0.1142

Visual inspection of the model's dense predictions against the original satellite imagery confirms that the network successfully learned spatial clustering and boundary detection purely from the 100 simulated point labels.