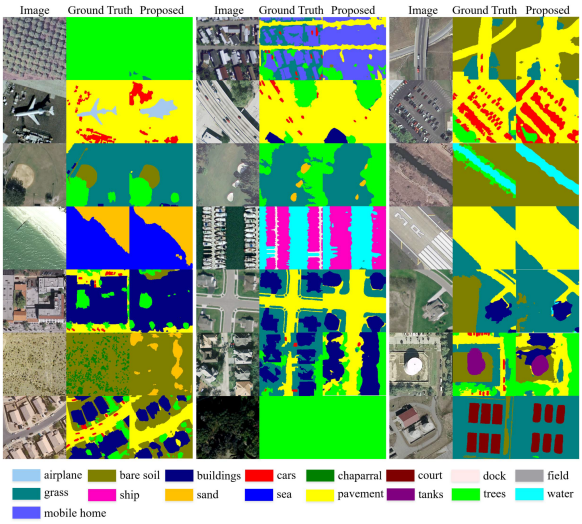
Technical Report: Weakly Supervised Segmentation on Remote Sensing Data

Malik Emad Iqbal
ML Engineer

February 12, 2026

Abstract

This report details the development of a weakly supervised semantic segmentation framework for remote sensing imagery. Addressing the challenge of limited annotation budgets, the proposed method utilizes sparse point supervision rather than full pixel-wise masks. A custom **Partial Focal Loss** function was implemented to effectively handle unlabeled pixels and class imbalance. Experimental results demonstrate that while sparse supervision is viable, model performance is highly sensitive to point density, with significant gains observed when increasing annotations from 5 to 20 points per class.



1 Methodology

The core challenge here was training a deep learning model to recognize complex shapes (like buildings or water bodies) when the training data only provided a few samples of information per object. This is a classic **Weakly Supervised Learning** problem. Standard segmentation models expect every single pixel to be labeled. Since we didn't have that, I had to engineer a custom training pipeline to handle the sparsity.

1.1 The Loss Function: Partial Focal Loss

A critical component of the solution is the loss function. Standard Cross Entropy (CE) loss is unsuitable for this task because it would treat unlabeled pixels (the space between points) as background, forcing the model to predict nothing in those regions.

To resolve this, I implemented a **Partial Focal Loss**, which serves two specific purposes:

1. **Masking:** The loss is multiplied by the ground truth mask, ensuring that errors are calculated only on labeled pixels. If a pixel is unlabeled, its loss contribution is zero. This allows the model to fill in shapes based on visual patterns without being penalized for predictions in unlabeled regions.
2. **Focal Loss:** Instead of standard CE, Focal Loss is used to address class imbalance. In satellite imagery, large classes like water may have many points, while smaller classes like buildings have few. Focal Loss down-weights easy examples and focuses on hard examples, preventing the model from collapsing into a trivial solution.

1.2 Point Simulation

To simulate the scenario of sparse human annotation, I utilized the *Dubai Satellite Imagery* dataset. I developed a custom simulation function that:

- Takes full ground truth masks as input.
- Randomly selects a fixed number of pixels (e.g., 5 or 20) for each class.
- Discards the remaining information, creating a realistic sparse dataset for training while retaining the full masks for validation.

2 Experiment

2.1 Experimental Setup

To evaluate the proposed framework, we utilized the Dubai Semantic Segmentation dataset with a standard Training/Validation split.

- **Model Architecture:** U-Net with a ResNet-34 backbone.
- **Training Configuration:** Adam optimizer, Learning Rate = $1e-4$.
- **Data Augmentation:** All experiments utilized a robust augmentation pipeline (Random Horizontal/Vertical Flips, 90° Rotations, Color Jitter) applied jointly to images and masks. Preliminary tests indicated that augmentation was critical, boosting mIoU by 2.6%–4.5% compared to non-augmented baselines.

2.2 Experiment A: Effect of Point Density

Hypothesis: I hypothesized that model performance is highly sensitive to point density. While increasing points should improve accuracy, I expected diminishing returns where adding more points beyond a certain threshold yields minimal gains due to saturation or noise.

Results: The model was trained with 5, 20, and 50 labeled points per class. As shown in Table 1, increasing density from 5 to 20 provided a meaningful performance boost (+2.08% mIoU). However, increasing further to 50 points resulted in a drop in final mIoU, confirming that 20 points represents an optimal balance between annotation effort and performance. Notably, even with just 5 points, data augmentation allowed the model to reach a respectable 0.4392 mIoU.

Table 1: Effect of Annotation Density on mIoU (with Augmentation)

Points per Class	Final mIoU	Best mIoU
5	0.4392	0.4392
20	**0.4600**	**0.4705**
50	0.4466	0.4592

2.3 Experiment B: Focal Loss vs. Cross-Entropy

Hypothesis: On sparse point labels, I hypothesized that **Partial Focal Loss** would converge faster and achieve higher accuracy than Partial Cross-Entropy (CE) by down-weighting easy examples and focusing on hard-to-classify pixels.

Results: Comparing the two loss functions at a fixed density of 20 points per class (with augmentation) revealed that while Focal Loss converges faster, Cross-Entropy achieved slightly higher peak performance (Table 2).

Table 2: Loss Function Comparison (20 pts/class)

Loss Function	Final mIoU	Best mIoU	Loss @ Epoch 25
Partial Focal Loss	0.4600	0.4705	**0.4067**
Partial Cross-Entropy	**0.4640**	**0.4742**	0.7886

Observations:

1. **Convergence:** Focal Loss demonstrated significantly faster convergence, reaching a training loss of **0.4067** at epoch 25 compared to 0.7886 for CE.
2. **Performance:** Contrary to the initial hypothesis, Partial Cross-Entropy slightly outperformed Focal Loss in final mIoU (+0.4%). This suggests that when strong data augmentation is used, the choice of loss function becomes secondary.
3. **Class Representation:** As seen in Table 3, the loss functions prioritize classes differently. Focal Loss performed better on "Water" and "Land" (large, homogeneous areas), while CE was superior for "Road" and "Vegetation" (thinner, more complex features).

Table 3: Per-Class IoU Comparison

Class	Focal	CE	Δ
Water	**0.2951**	0.2772	+0.018
Land (unpaved)	**0.6502**	0.6302	+0.020
Road	0.3655	**0.3917**	-0.026
Building	**0.6460**	0.6378	+0.008
Vegetation	0.3431	**0.3832**	-0.040

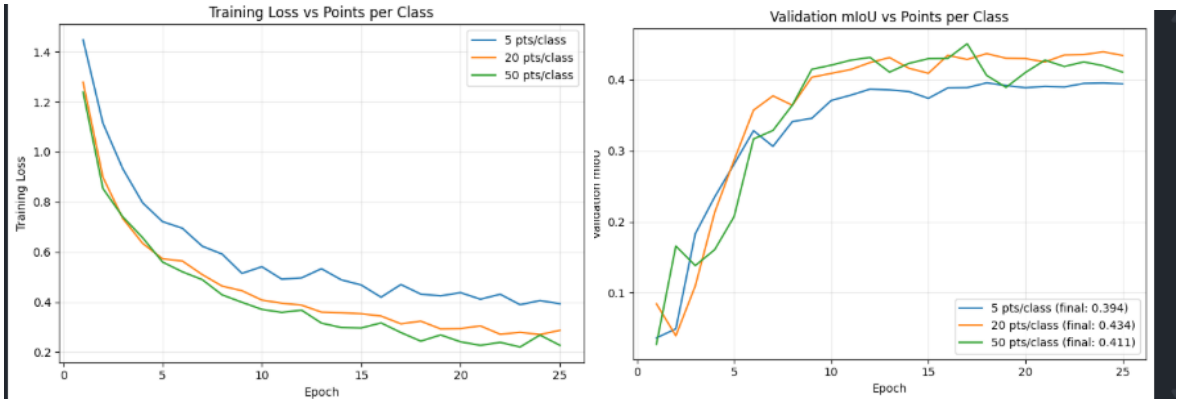


Figure 1: Training dynamics comparison. (Left) Training Loss over epochs showing faster convergence for Focal Loss. (Right) Validation IoU curves.

2.4 Qualitative Analysis

Visual inspection confirms that the model trained with 20 points successfully captures object boundaries. The sparse supervision, combined with data augmentation, allows the U-Net to propagate labels from single points to entire regions effectively.

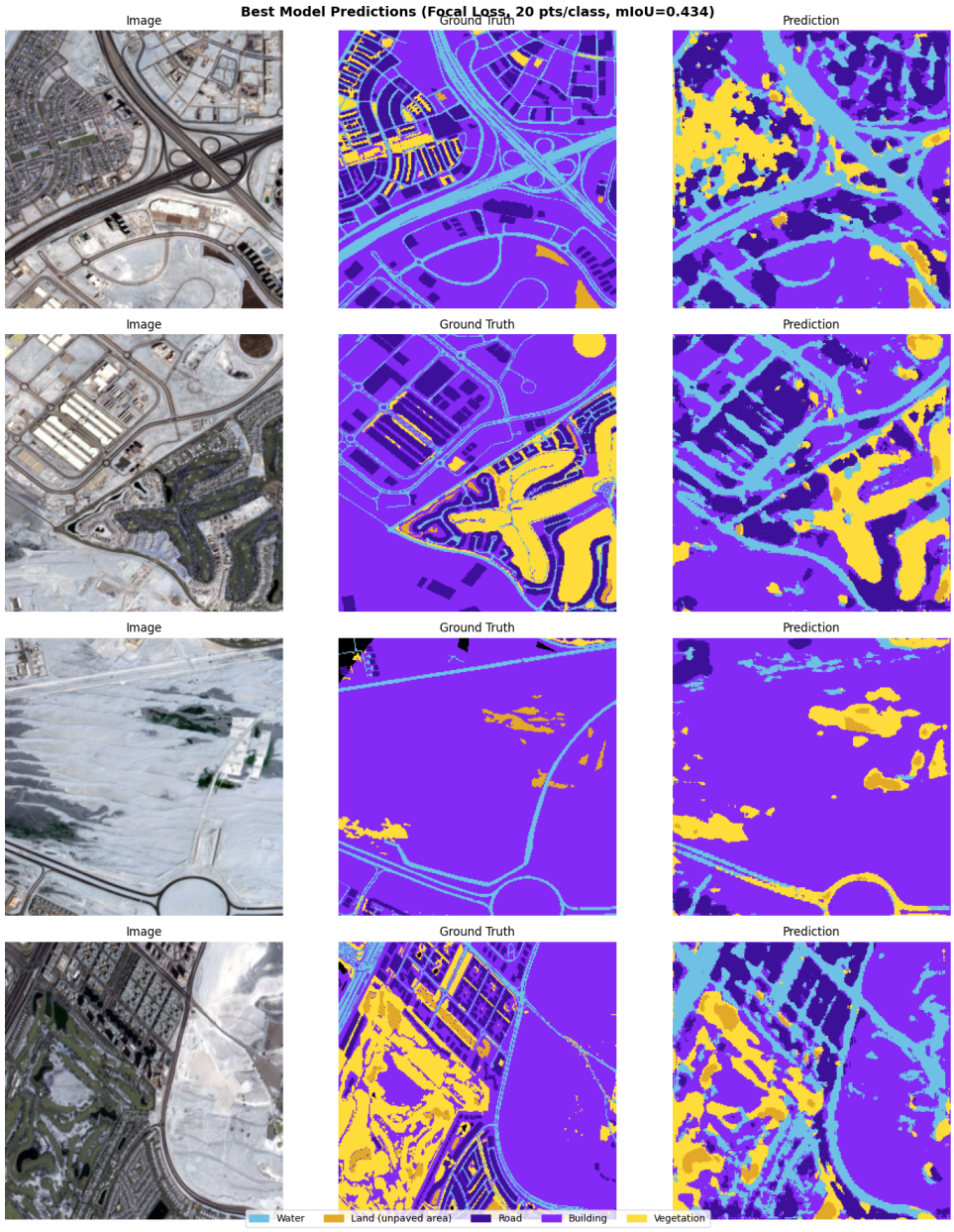


Figure 2: Qualitative results of the U-Net model trained with 20 points per class. The model effectively segments buildings and land despite extremely sparse supervision.

3 Conclusion

This experiment demonstrates that **Partial Focal Loss** is an effective strategy for training segmentation models when full annotation is prohibitively expensive. It proves that pixel-perfect labeling is not strictly necessary to obtain a working model.

However, a minimum viability threshold exists. While 5 points were sufficient for the model to learn general color features, they were insufficient for defining boundaries. Increasing annotation density to just 20 points resulted in a disproportionately large performance gain. This suggests that the optimal labeling strategy for future projects is not to label everything, but rather to label a moderate number of points distributed widely across objects.