Microsoft | AI for Good Lab

Data-Centric Land Cover Classification Challenge

Akram Zaytar*, Simone Fobi*, Caleb Robinson, Anthony Ortiz

* Equal Contribution

The 34th British Machine Vision Conference


Drone
Spatial Resolution: ~ 5 cm/px
Temporal Cadence: As needed


Sentinel 2A
Spatial Resolution: ~ 10 m/px
Temporal Cadence: ~ weekly


Landsat Collection
Spatial Resolution: ~ 30m/px
Temporal Cadence: ~ 2 weeks

Data Centric Approach to Building Damage Assessment

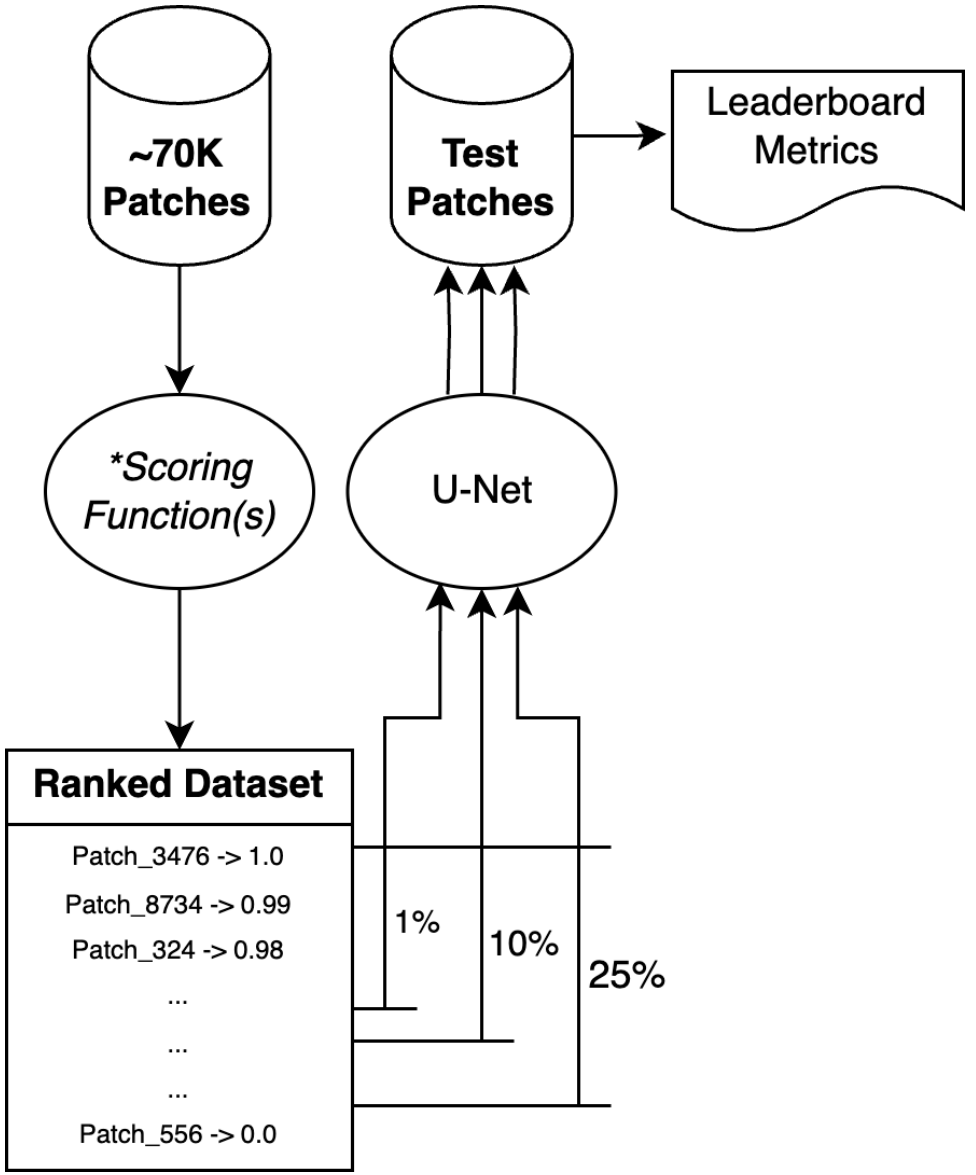
Extent of a disaster

Set of patches

Given a disaster event that covers a large area, which patches should be labelled for high performance and rapid response ?

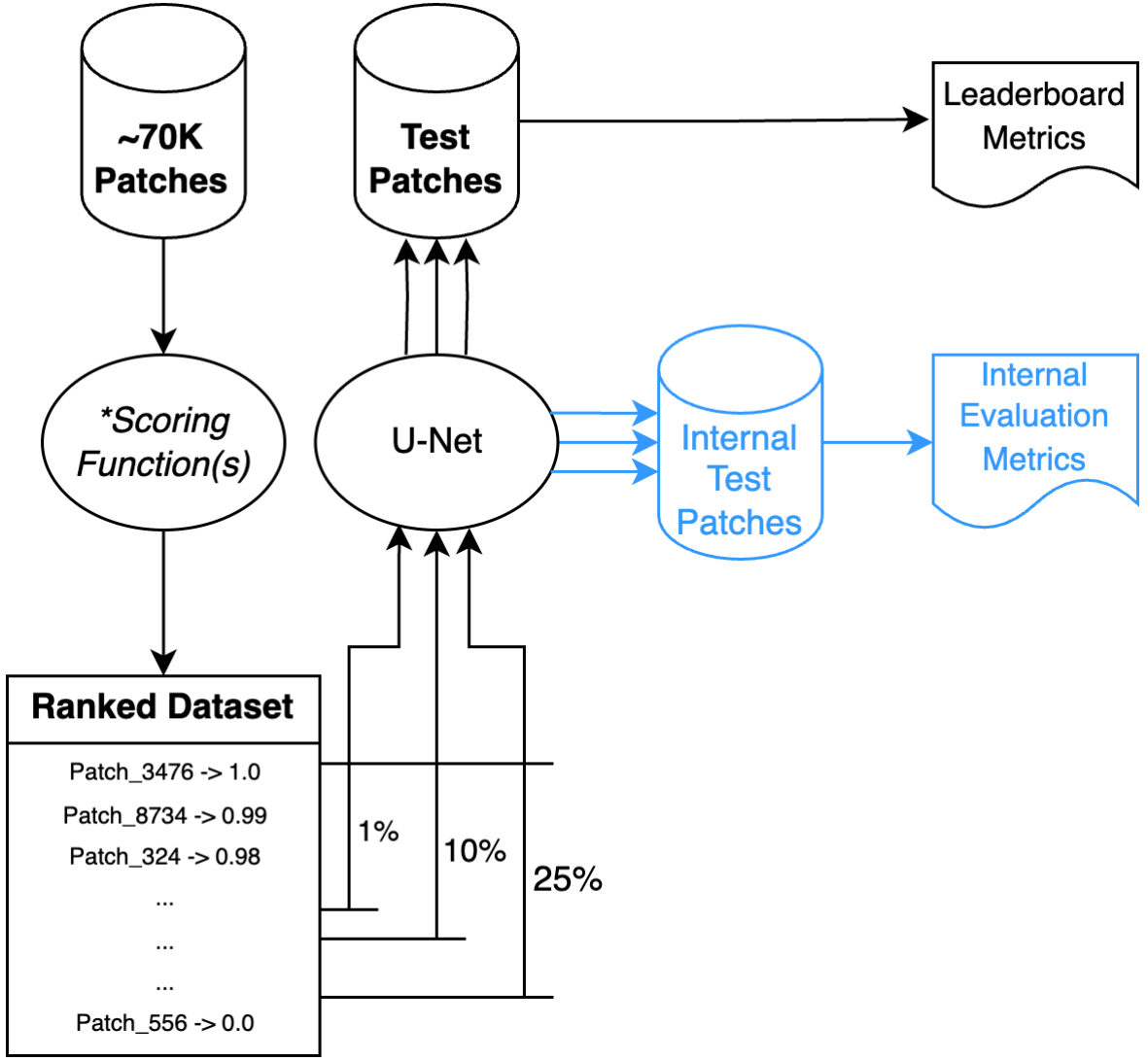
Challenge Overview

- **Motivation:** Efficiently selecting where to label and identifying valuable samples is crucial due to limited resources for labeling & training.
- **Goal:** Prioritize high-quality data samples over quantity for training semantic segmentation models. Develop ranking strategies to score data samples between 0 to 1.
- **Evaluation Pipeline:** Use the scores to train U-Nets with the top 1%, 10%, and 25% data samples. Leaderboard scores are based on average Jaccard of these models on a held-out test set.



Internal Experimentation Setup

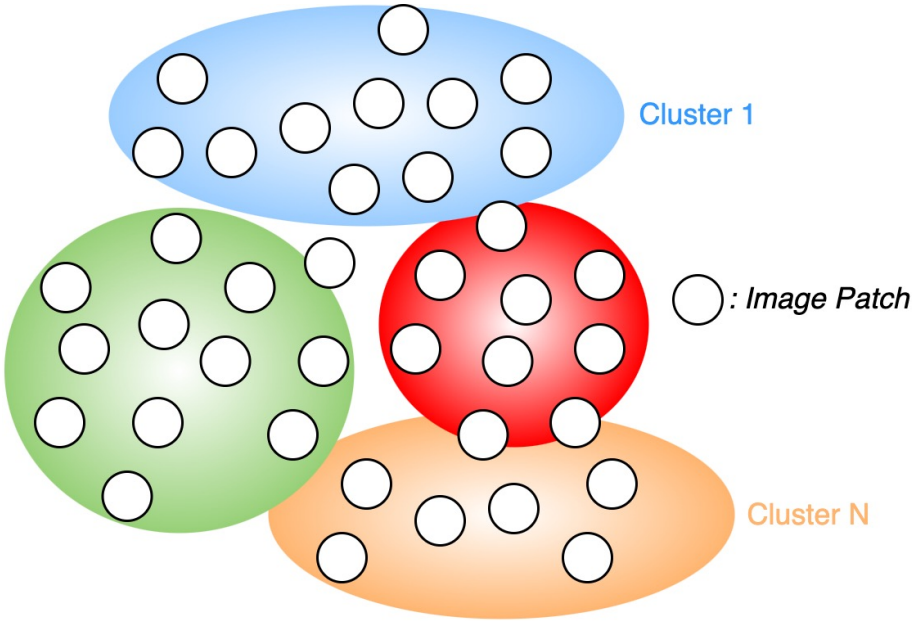
- **Data Split:** Randomly divided the competition validation set into two equal halves by tile ID, creating our internal “Validation” and “Test” sets.
- **Censored Unlabelled Samples:** Removed samples with no data labels (0) & cloud coverage (15).
- **Training Setup:** Focused primarily on training with 1% and 10 % sample sizes to speed up experiments. No early stopping.
- **Performance Uncertainty:** Conducted 3 training runs using different seeds to assess metric uncertainty on the test set.
- **Consistent Metrics:** Followed the same metric calculation method as specified in the competition's GitHub repository.
- ***We want to experiment with different scoring functions and compare the results.***



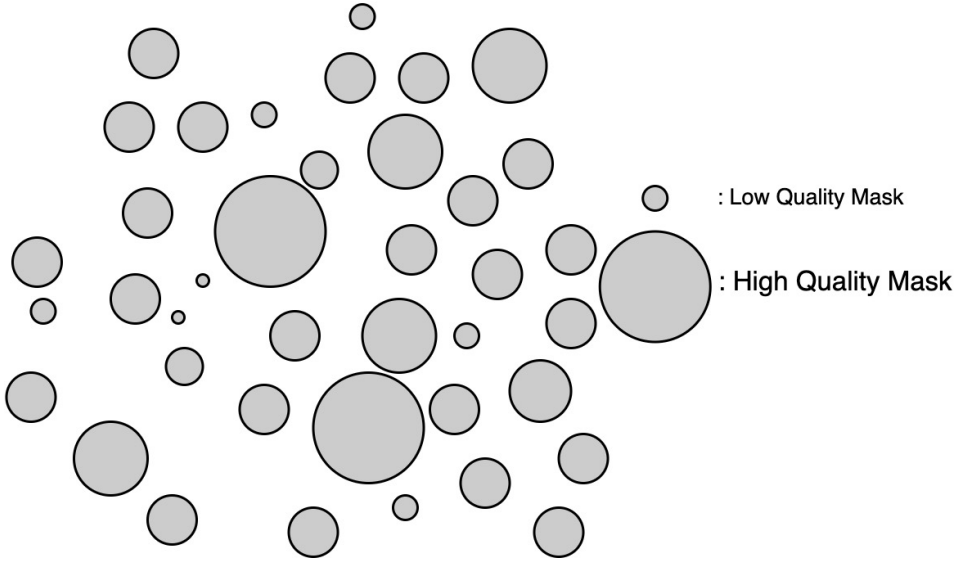
Scoring Functions Overview

We explored 3 sets of scoring functions:

- **Image-based methods:** focus on the RGB image regardless of the corresponding masks.
- **Mask-based methods**: focus on the single-channel mask regardless of the corresponding RGB or DEM images.
- **Hybrid methods**: focus on combining the image & mask-based function scores.



Example: Image-based Clustering.



Example: Each mask is scored by quality.

Submission 1: Vendi Clustering

Ecological diversity: effective # of species = exponential of entropy

$$VS_k(x_1, \ldots, x_n) = \exp\left(-\operatorname{tr}\left(\frac{\boldsymbol{K}}{n}\log\frac{\boldsymbol{K}}{n}\right)\right)$$

*Vendi score as the Von Neuman Entropy for similarity matrix (**K**).*

Scoring Function: Dynamic clustering with Vendi Scores

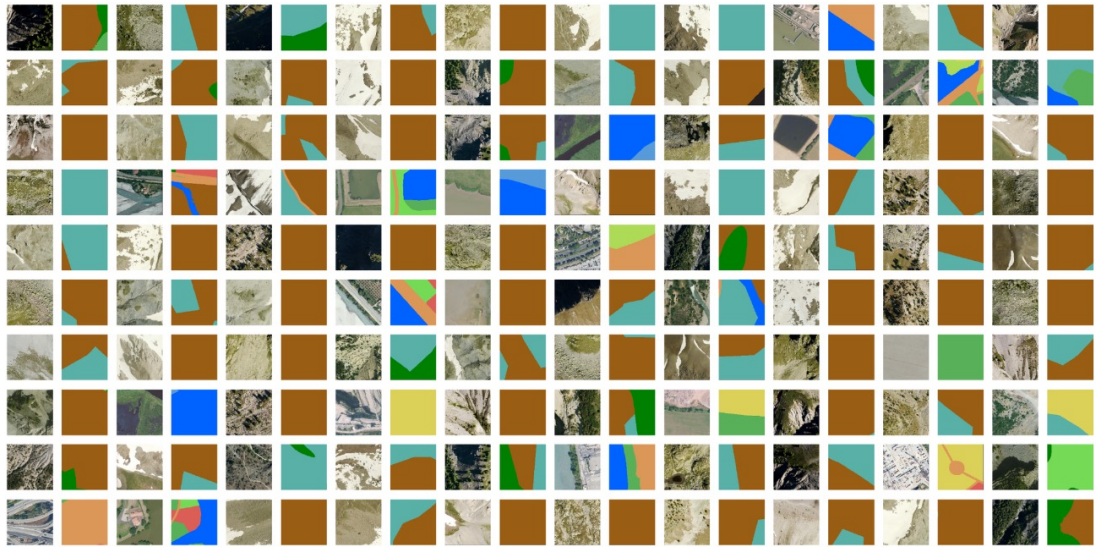
1. For each RGB patch, extract its features using ViT/ResNet encoder.
2. Apply KMeans clustering to the features.
3. Calculate the mean Vendi score [*Friedman et al*] for the current cluster arrangement.
4. Determine the change in the mean Vendi score compared to the previous iteration.
5. Continue this process until the average change in the mean Vendi score over the last few iterations is below a small threshold, indicating stability in the clustering.

Ranking procedure: Rank the samples by sequentially sampling from each cluster. This method ensures a diverse representation from all identified clusters in the final sample set.

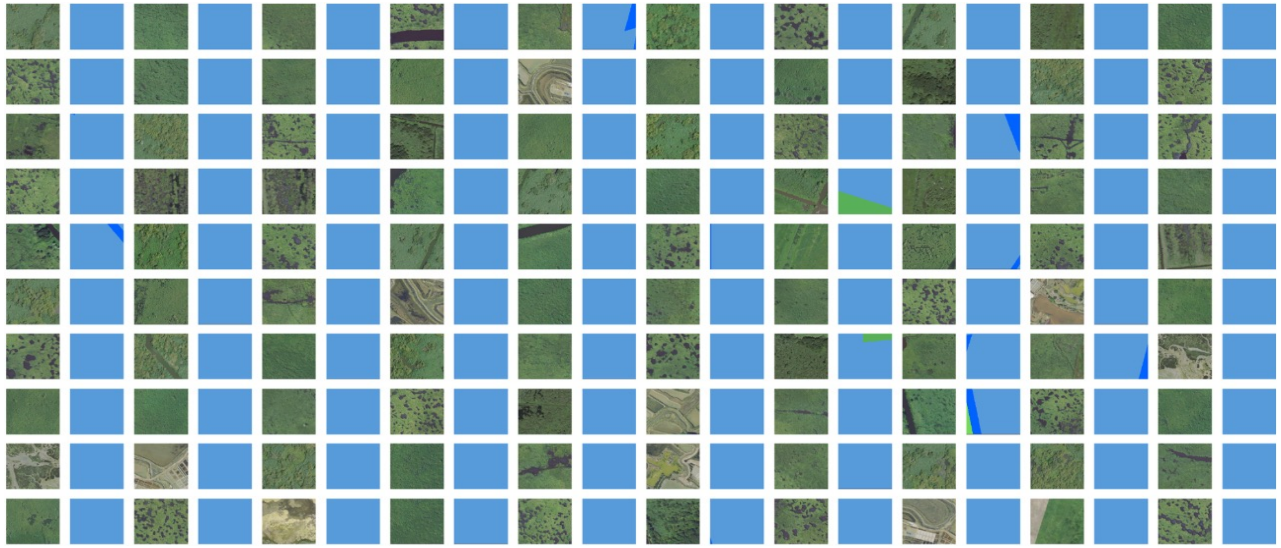
Image-Mask Consistency across Different Vendi Clusters

Cluster 1

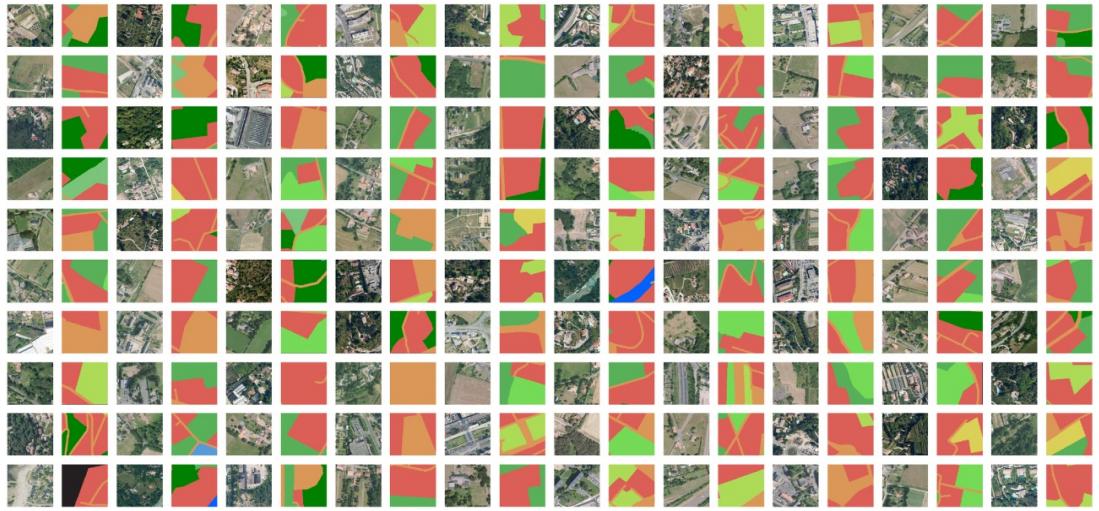
Patch Mask



Cluster 2

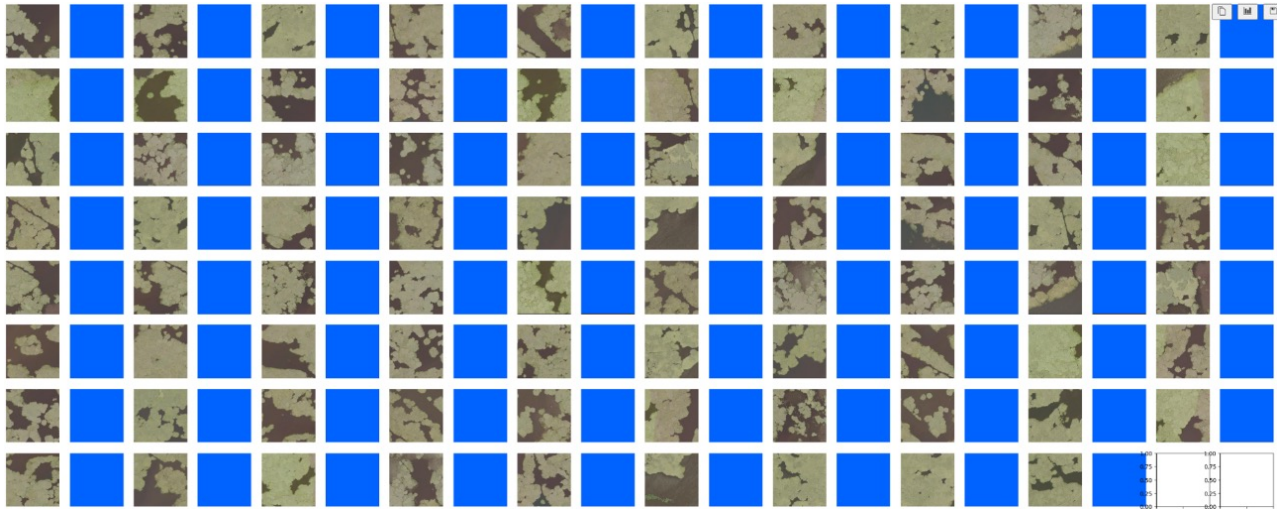


Cluster 3



...

Cluster N



Submission 2: Diversity + Entropy Sampling

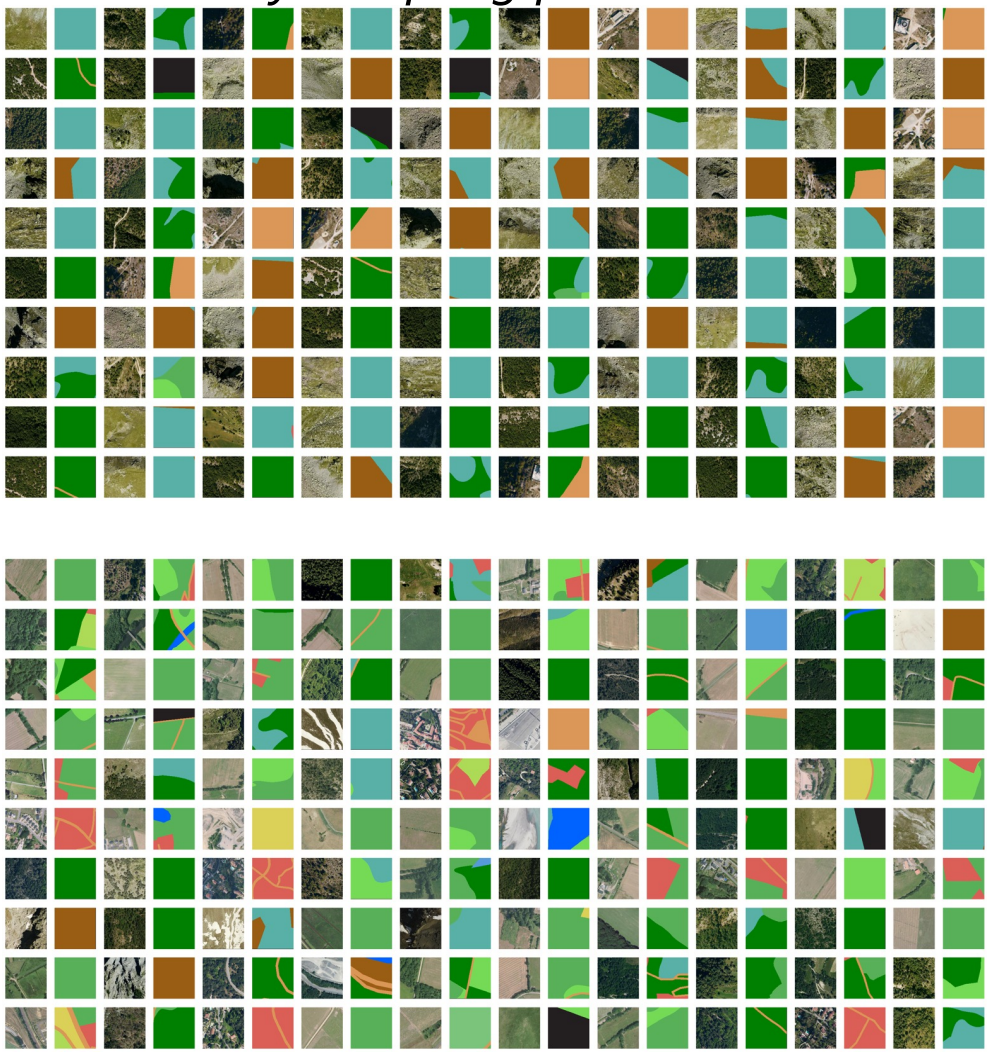
Average scores from diversity and entropy sampling

Diversity Sampling: Compute L2 Norm of image representations and sample across percentile bins.

Entropy Sampling

1. Train a model on an initial sample set.
2. Use the model's probabilistic predictions to evaluate each sample.
3. Samples are scored based on the average patch-wise entropy, where higher entropy indicates greater uncertainty and potentially higher value for training.
4. The rationale is that samples with higher entropy (uncertainty in model predictions) are more informative and should be prioritized for training.

Diversity sampling pseudo-clusters



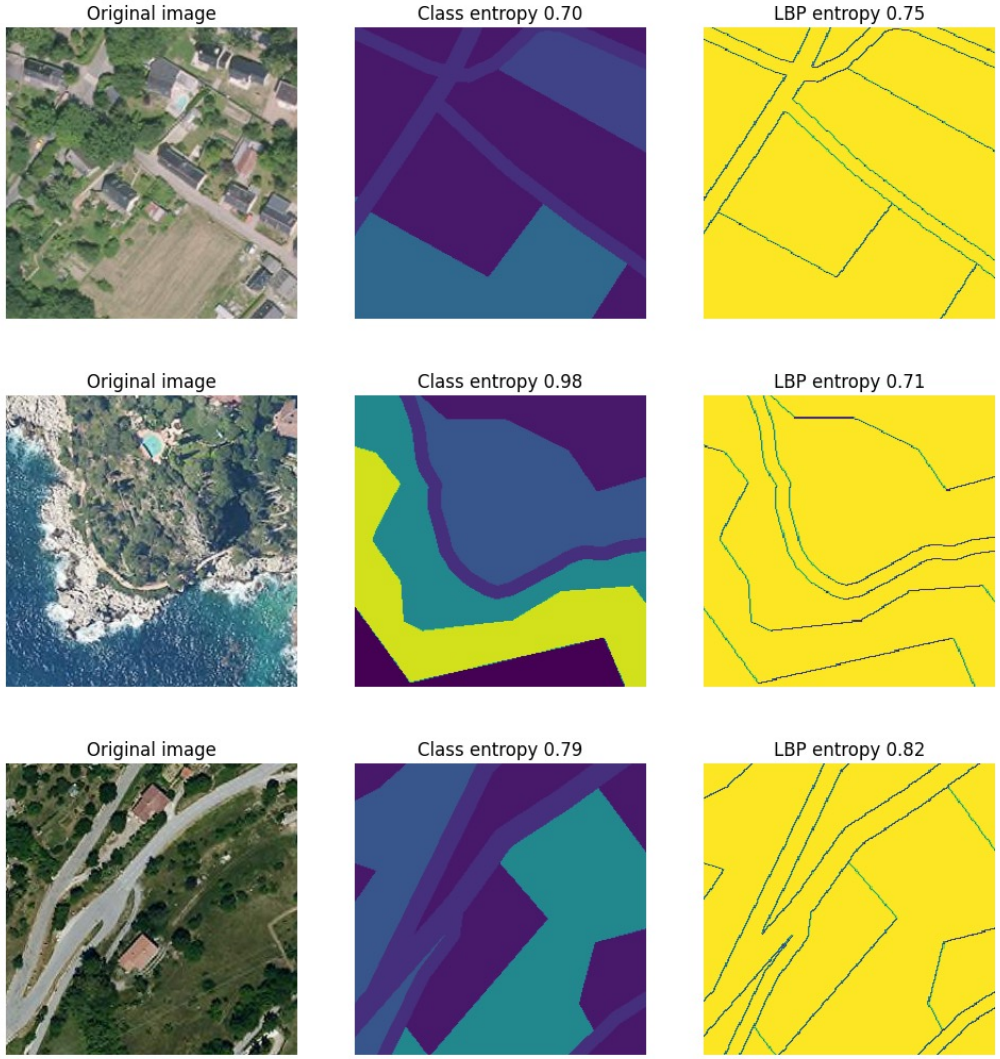
Submission 3: Mask Entropy & Local Binary Pattern

Calculates sampling weights using mask & LBP entropy:

Entropy Calculation: uses mask unique class ratios.

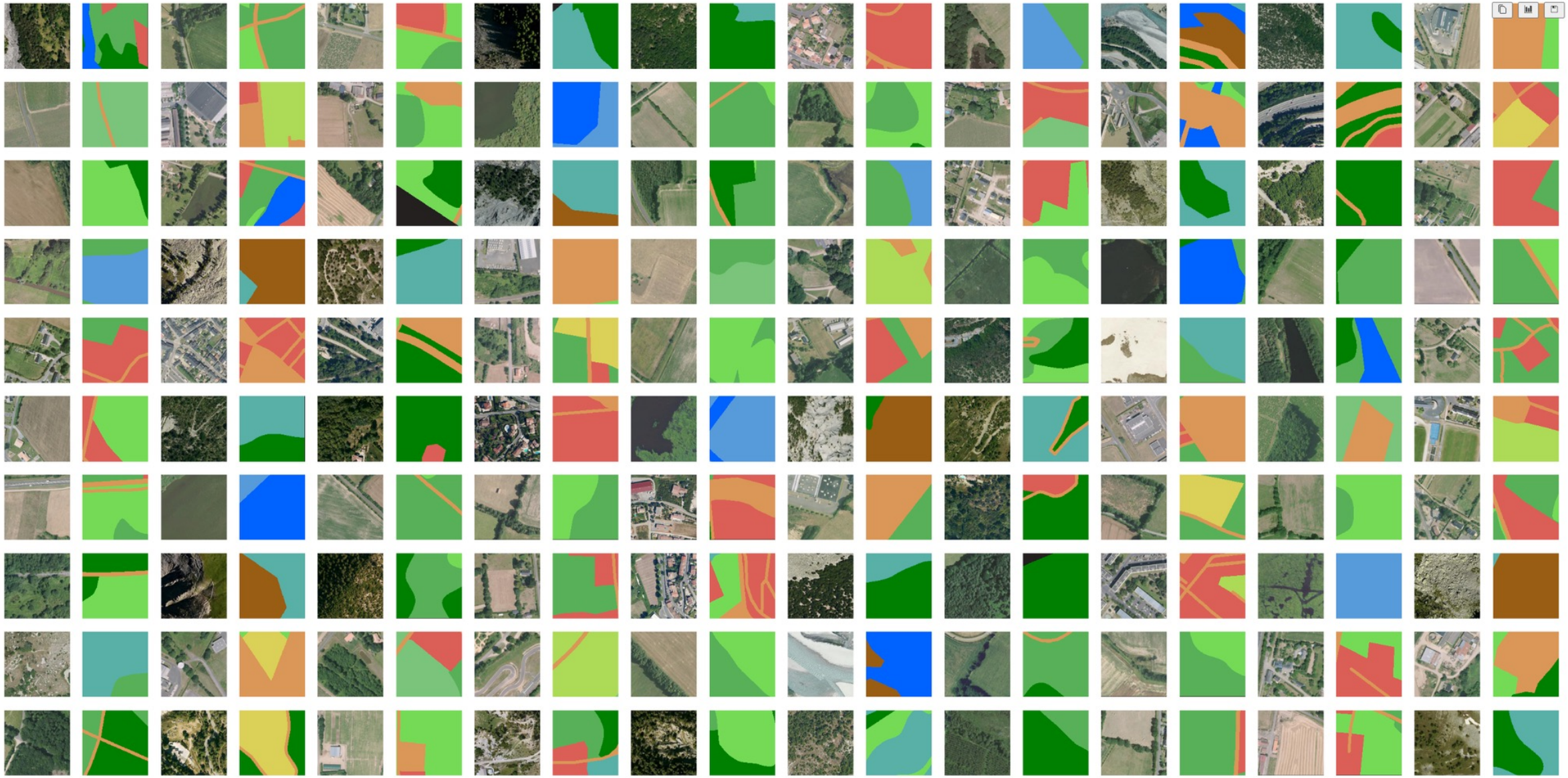
Local Binary Pattern (LBP): Mask texture entropy.

1. For each pixel, compare its value with its 8 neighbours, encoding these comparisons as a binary number.
2. Generate a new binary image, capturing the texture pattern of the mask.
3. Calculate the binary entropies for each mask.



Mask Complexity Ranking (top 100)

Patch→Mask



Submission 4: Vendi Clustering + Mask Complexity

Observation

Patch diversity ranking is more effective than mask complexity in low regimes.

Method

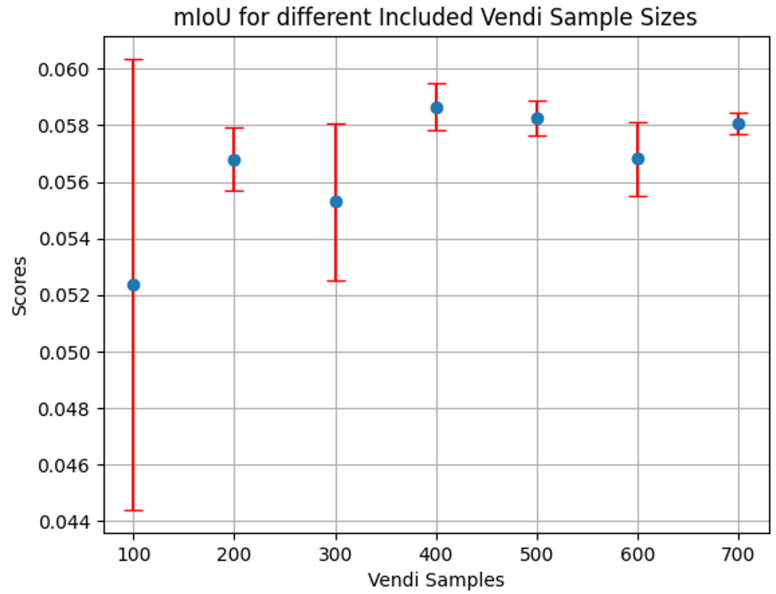
A two-phase ranking strategy.

1. Use Vendi Clustering to rank the first set of data samples.
2. Remaining samples are ranked based on their mask complexity scores.

We leverage both diversity and complexity in the ranking process.

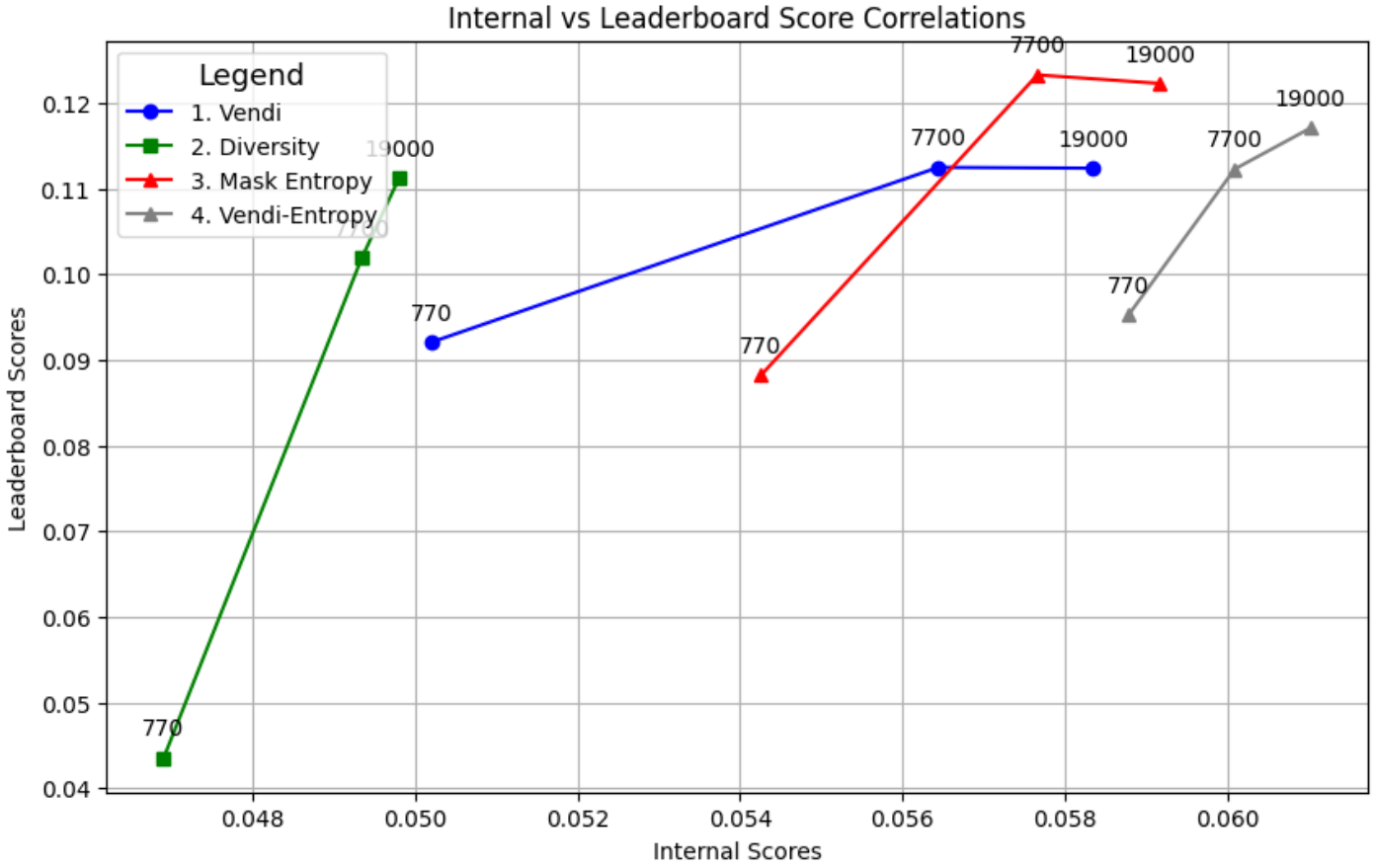
Submission Approach

The top 700 patches were selected based on their Vendi Clustering scores, emphasizing diversity. The subsequent samples were chosen based on mask complexity scores.



Internal score variance for different thresholds of included Vendi samples (remainder until 770 use mask complexity ranking).

Results & Learnings



Hybridization Insights

- Mean and ranking-based hybridization did not yield superior results.
- Future directions: probabilistic sampling & threshold-based approaches.

Mask Complexity and Quality

- Focusing on mask complexity and quality tended to produce better outcomes.
- Indicates the significance of these factors when making use of weak labels.

What did not Work?

Active learning

- uncertainty sampling
- query by committee
- Expected model change
- min-margin sampling.

Ranked Sampling within Clusters

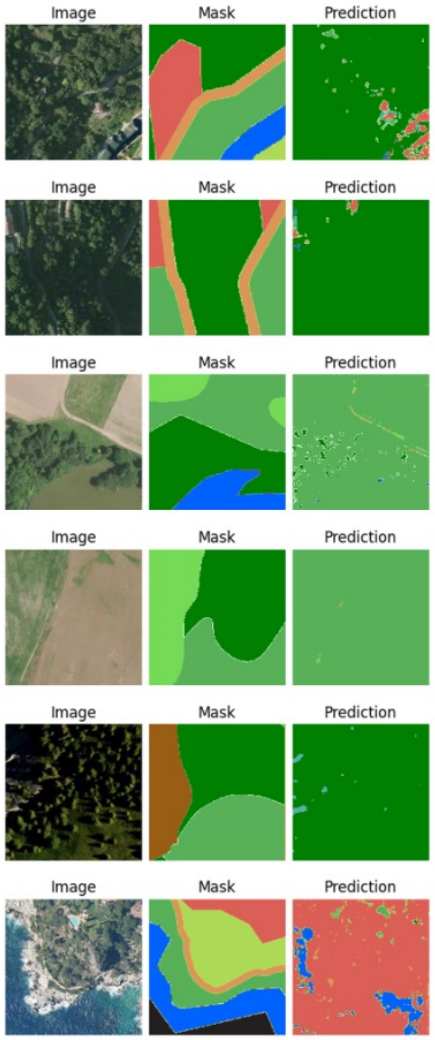
Filtering out "low-effort labels": such samples are already panelized by mask complexity ranking.

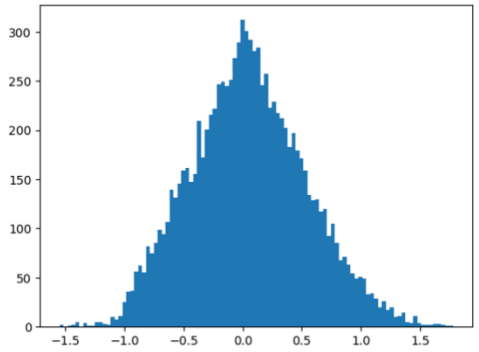
Variance(Encoder[augmentations]): same reason as above.

To be studied separately:

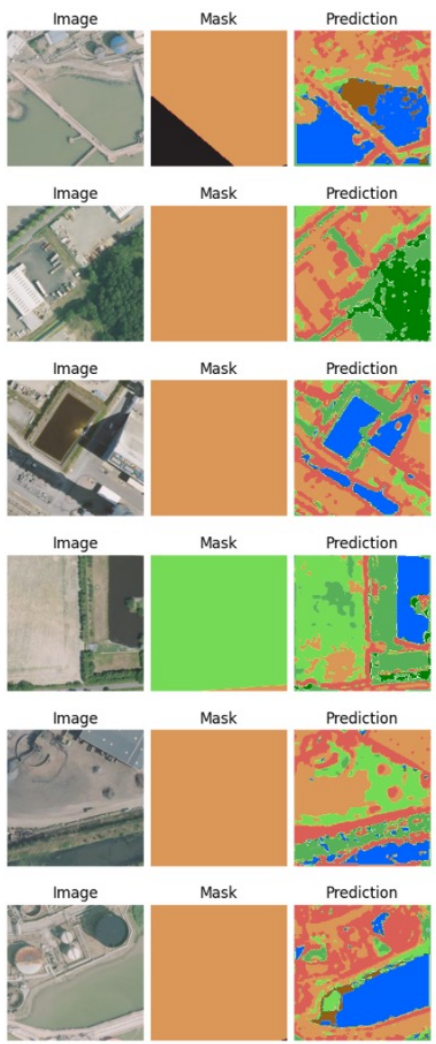
- *Low regimes != High regimes*
- *Training with augmentation != without augmentation.*

Insight: Filter out “Low annotation effort” Masks





*Note: Effort can be negative if the model is unable to predict the details.



Method:

1. Train an FCN on the dataset.
2. Use the model to predict masks on the data.
3. Measure:

$$effort = complexity(predicted\ mask) - complexity(label)$$

Insight: our dataset is the augmented one.

Procedure

1. Generate all possible augmentations of each sample.
2. Get the representations of the patches while keeping track of their origins.
3. Measure patch variance using its representations.
4. Average-Aggregate the augmentations scores to get the patch global score.

High-Variance

Low-Variance

Future Directions

Complexity Ranking without Labels

- **Early Phase**: Initially employ Vendi clustering to select the first N samples for labeling.
- **Later Phase**: Use a robust, non-overfitting model (like a Fully Convolutional Network) to predict labels for the remaining samples.
- **Entropy Measurement**: Calculate label entropy on these predicted patches.
- **Ranking**: Rank the remaining inputs based on the prediction entropy.
- **Batch Addition**: Continuously integrate new batches using this methodology.

Label-Correctness Algorithms for weakly supervised datasets

- Plan to explore algorithms focused on verifying and enhancing the correctness of labels in training datasets.

Active Learning Method Trade-offs

- Study the effectiveness and trade-offs of various active learning methods in different dataset sizes (low vs. high size regimes).

Geospatial Dataset Benchmarking

- Aim to benchmark the current methods across other geospatial datasets to assess their generalizability and effectiveness.

Training Procedure Effects

- Investigate how different training procedures, like data augmentation and semi-supervised learning, impact the benefits derived from subset selection.

Hybridization Methods

- Explore advanced hybridization techniques, such as weighted sampling using probabilistic distributions and threshold-based sampling, for more effective data scoring and selection.

References

[1] Friedman, Dan, and Adji Bousso Dieng. "The vendi score: A diversity evaluation metric for machine learning." *arXiv preprint arXiv:2210.02410* (2022).