Large Scale Multi-Label Image Classification of Fine-Grained Wetland Vegetation in the Okavango Delta

By Jason Mai and Hitha Varganti

Berkeley UNIVERSITY OF CALIFORNIA

Introduction

The UC Berkeley Environmental Systems Dynamics Lab is working on a project that involves using LiDAR and remotely-sensed data to inform land-cover mapping of the Okavango Delta in Botswana, to explore questions about hydrology, ecology, and geomorphology. The National Geographic Society has collected over ten-thousand geolocated 360° images to be labeled, in order to train and validate a classifier that distinguishes the delta's fine-grained vegetation types. An automated multi-label image classifier will (i) accelerate annotation of the remaining images, and (ii) yield a validation set for the satellite- and LiDAR-driven land-cover mapping.

Data and Methods

- Over 10,000 original 360° rectangular panoramic 5952x2976 images taken from boats traversing channels in the Okavango in 2023.
- We split images in half creating 2 benefits: Centers ROI & creates perfect 2976x2976 square images for neural networks.

Annotation and Labeling

- Used LabelBox software to upload and label initial images (~600)
- Classifications:
  - Level 1:
    - Open Water
    - Floating Vegetation
    - Reedlands and Permanent Swamps
    - Seasonal Tall (sedge) Wetlands
    - Seasonal Short (grass) floodplains
    - Riparian Forest
    - Broadleaf Woodland
    - Open Woodland
    - Bare Ground
  - Level 2:
    - Water Lilies
    - Rafts
    - Salvinia Molesta
    - Cyperus Papyrus
    - Phragmites
    - Miscanthus Junceus
    - Vossia Cuspidata
- Level 1 Classifications also received "Distance" label categorizing Foreground, Nearest Bank, or Background.

Sampling

- Stratified Sampling: Keep same proportion of each label for training/validation/test splits.
- Group-Aware Sampling: Keep left/right halves together between splits to prevent data leakage as left/right halves of a given image share context.

Model Architecture

Pretrained on ImageNet-22k and fine-tuned on ImageNet-1k models:

- Swin Transformer v2 (hierarchical attention)
- ConvNeXt (CNN-based modern architecture)

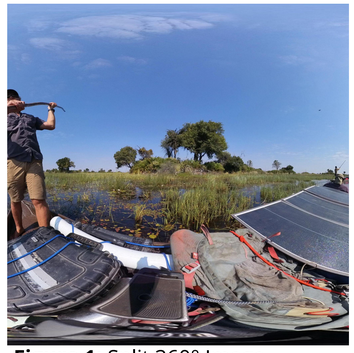

Figure 1: Split 360° Image Example

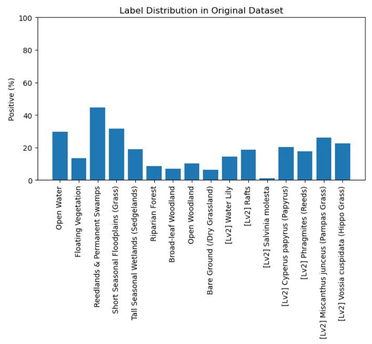

Figure 2: Distribution of Image Labels in Entire Dataset

Image Augmentation

Prevent overfitting and enhance generalization with:

- Horizontal roll (p=0.5), affine (±5° rotation/scale/shear, ±5% translation), resized crop (80–100% area), flips (horizontal p=0.5, vertical p=0.1), color jitter (brightness/contrast/saturation ±30%, hue ±0.1), gaussian blur (p=0.15), random erasing (2–10% area, p=0.25), MixUp (α=0.4, p=0.3), CutMix (α=1.0, p=0.3)

Asymmetric Loss (ASL)

- Suppresses gradients from easy negatives (common in rare classes)
- Focuses training on hard positives/negatives.
- Clips extreme predictions to avoid overconfidence.

Class-Specific Residual Attention (CSRA)

- Attention to image regions most relevant to each label

ML-Decoder

- Generates class-specific feature vectors by attending to critical image regions.
- Learns label correlations via all queries training jointly.

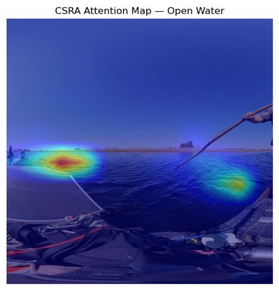

Figure 3: CSRA Attention Map

Evaluation Strategy

- Threshold tuning: Optimize per-class thresholds on validation data to maximize F1.
- Metrics:
  - Average Precision (AP): Average the precisions of all possible recalls
  - Precision: % correct positive predictions
  - Recall: % of true positives correctly predicted
  - F1: Harmonic mean of precision/recall

Results

Collectively, these results indicate that although advanced techniques like ASL, CSRA, and ML-Decoder can deliver slight performance enhancements when meticulously optimized, the primary and most substantial gains in classification performance are more likely to come from focusing on expanding the quantity and quality of the training data.

Swin V2 B (384 px, 88 M params)

Method	mAP
BCE (Linear head, n=319)	0.511
BCE (Linear head, n=605)	**0.583**
ASL (Linear head, γ_-=4, clip=0.05)	0.568
ASL (Linear head, γ_-=4, clip=0.02)	**0.583**
CSRA (H=1, λ=0.1)	0.537
CSRA (H=1, λ=0.3)	0.572
CSRA (H=8, λ=0.1)	0.535
CSRA (H=8, λ=1.0)	0.481
ML-Decoder	0.506

Figure 4: SwinV2 Results

ConvNeXt B (384 px, 88 M params)

Method	mAP
BCE (Linear head)	0.543
ASL (Linear head, γ_-=4, clip=0.05)	0.533
ASL (Linear head, γ_-=4, clip=0.02)	**0.559**
CSRA (H=1, λ=0.1)	0.523
CSRA (H=1, λ=0.4)	0.544
CSRA (H=6, λ=0.1)	0.536
CSRA (H=6, λ=0.4)	0.485
ML-Decoder	0.546

Figure 5: ConvNeXt Results Mirrored settings used in original papers

Class	mAP
Open Water	0.961
Reedlands & Permanent Swamps	0.807
Broad-leaf Woodland	0.218
Open Woodland	0.335
Cyperus papyrus (Papyrus)	0.919

Figure 6: Selected per-class mAP from best performing SwinV2

Discussion

Challenges

- Sample size, noisy image labels, occasional label errors.
- Severe class imbalance - rare yet ecologically significant classes have fewer training examples, thus complicating model training and threshold selection.

Further Research

- Use Stratified Group K-Fold Cross-Validation to increase threshold robustness.
- Collapse similar classes like Woodlands + ignore "Background" labels to reduce noise.
- Explore accuracy of using Segmentation models.

References and Acknowledgements

Ben-Baruch, Emanuel, et al. "Asymmetric Loss for Multi-Label Classification." *arXiv* preprint arXiv:2009.14119 v4, 29 July 2021, https://arxiv.org/abs/2009.14119.

Ridnik, Tal, et al. "ML-Decoder: Scalable and Versatile Classification Head." *arXiv* preprint arXiv:2111.12933 v2, 31 Dec. 2021, https://arxiv.org/abs/2111.12933.

Zhu, Ke, and Jianxin Wu. "Residual Attention: A Simple but Effective Method for Multi-Label Recognition." *arXiv* preprint arXiv:2108.02456 v2, 19 Aug. 2021, https://arxiv.org/abs/2108.02456.

Thank you to Dr. Lukas WinklerPrins, Dr. Laurel Larsen, Magali Ruer, Winnie Yang, Alexander Brown, Meixian Li, Sophia Meyers, Octavia Crompton, David Pham, the ESDL Team, US Army Corps of Engineers, National Geographic Okavango Wilderness Project, and Wild Bird Trust.