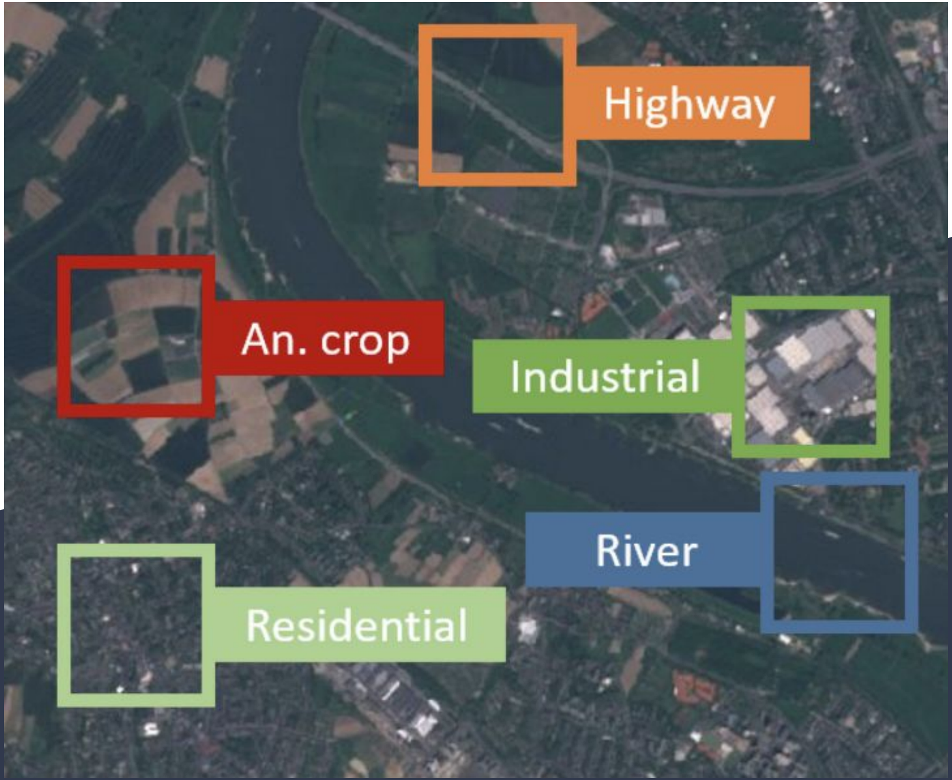
Land Use and Land Cover Classification of Satellite Imagery

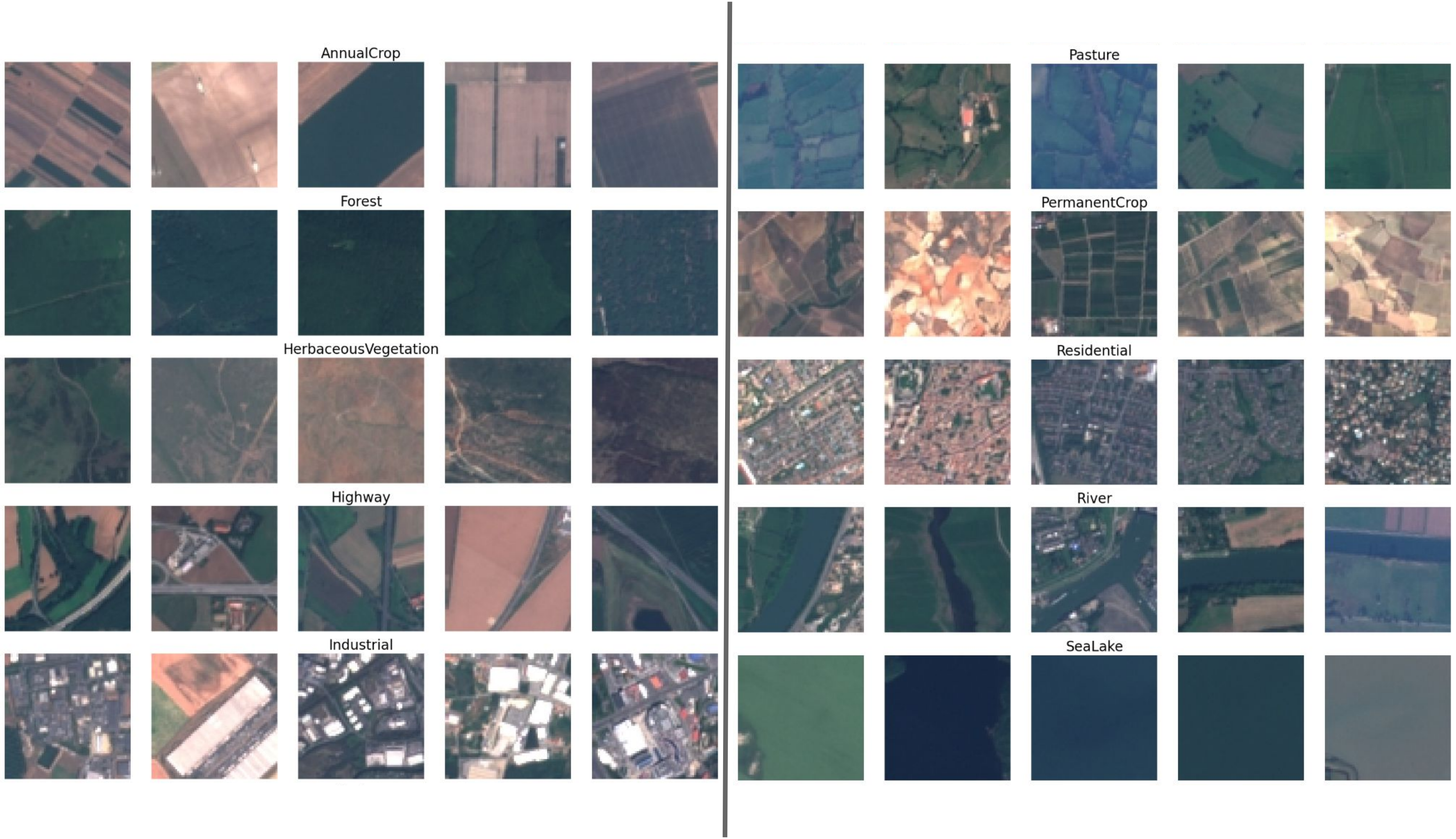
Andrew Loeber, Yeshwanth Somu, Zhifei Dong



Dataset Information

- Data comes from **EuroSAT** dataset
- **27,000** satellite images of land cover in JPEG format
- Each image is a **64x64 RGB** pixel grid
- Labeled into **10 classes**, shown on the right
- Split into train/validation/test in **70%/15%/15%**, stratified by class label

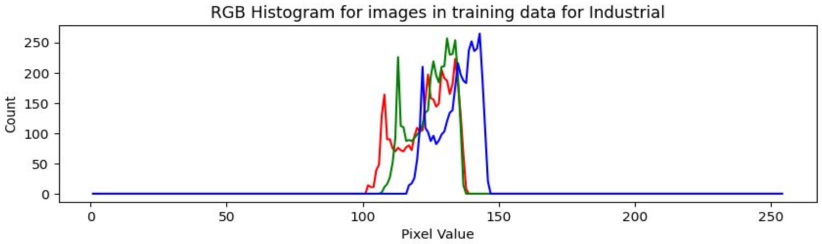
Class ID	Class Label	Number of Images
0	Industrial	2500
1	Residential	3000
2	Highway	2500
3	Annual Crop	3000
4	Permanent Crop	2500
5	Pasture	2000
6	Herbaceous Vegetation	3000
7	Forest	3000
8	River	2500
9	Sea Lake	3000

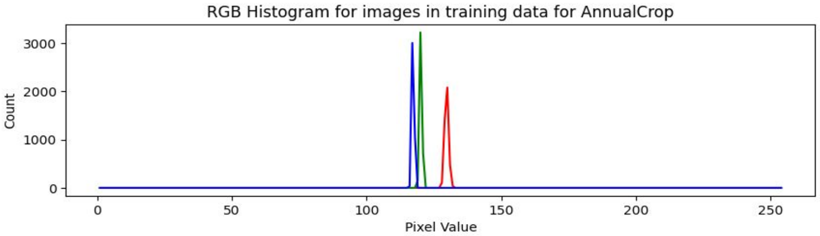

AnnualCrop
Forest
HerbaceousVegetation
Highway
Industrial
Pasture
PermanentCrop
Residential
River
SeaLake

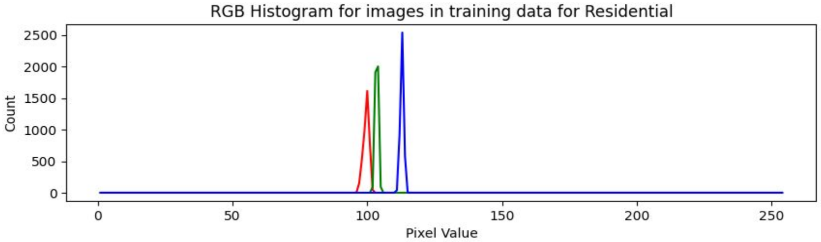
Input Features

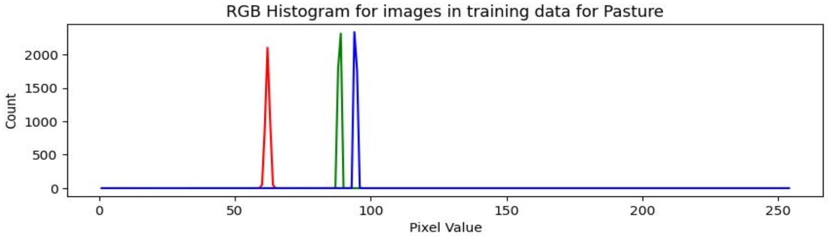
Color	Average RGB Values
	K-Means Color Bin Histograms
Texture	Gray-Level Co-occurrence Matrix (GLCM)
Complex	EfficientNet Embeddings

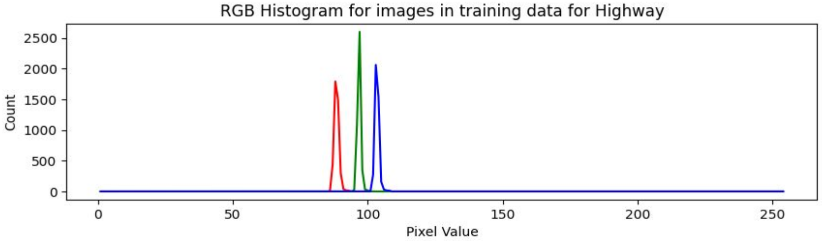
Average RGB Values

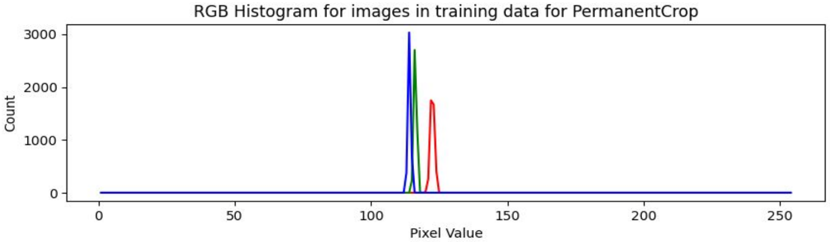




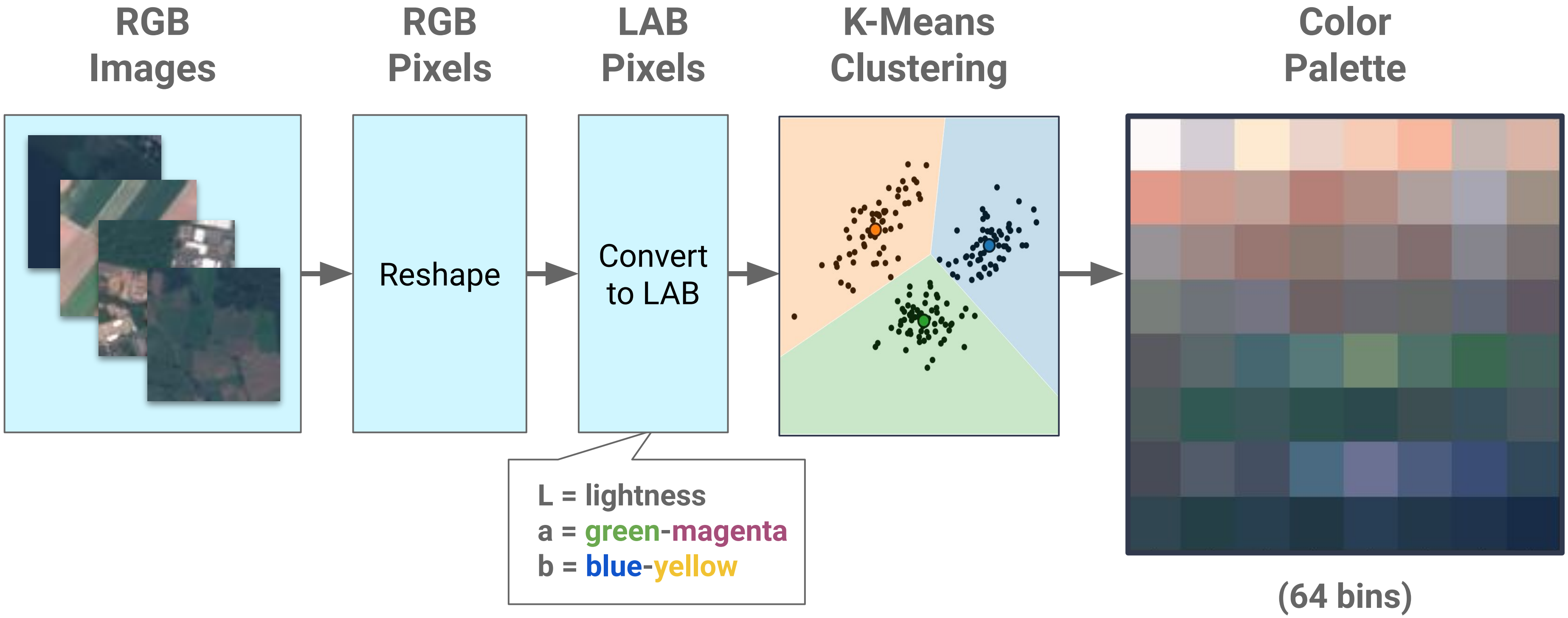








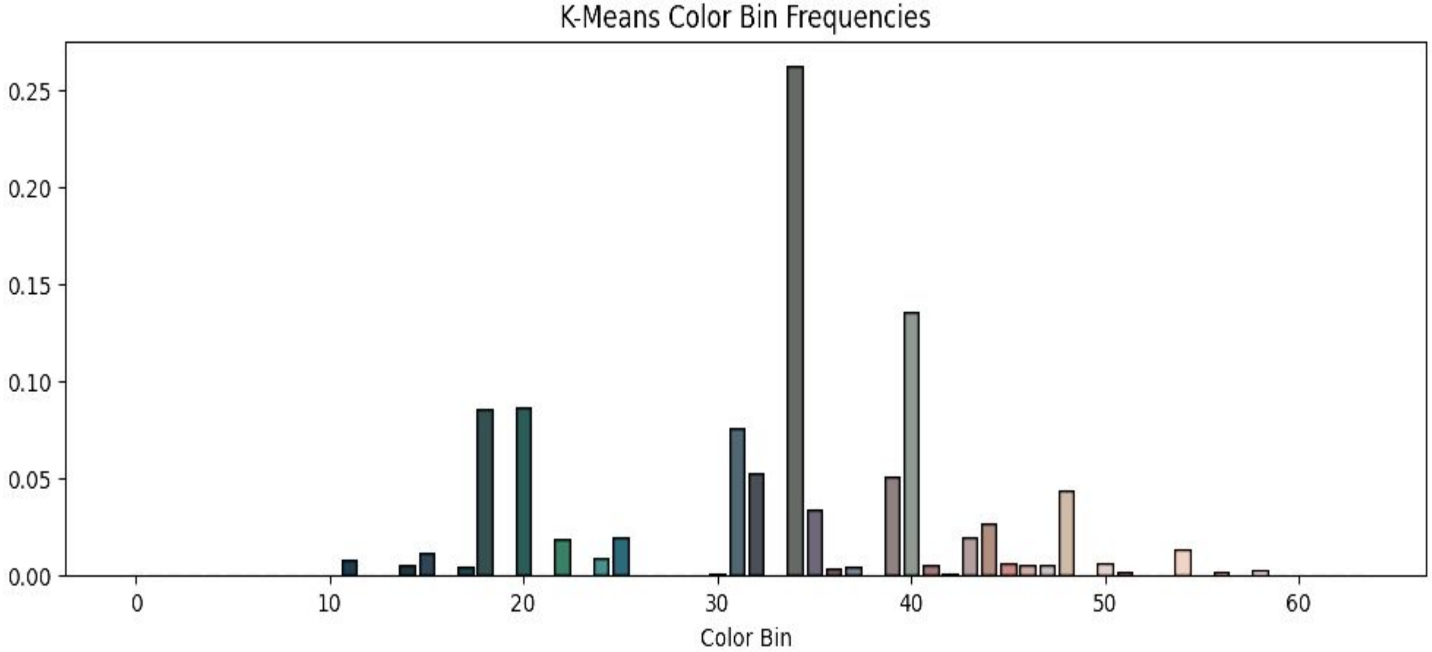
K-Means Color Bin Histograms



K-Means Color Bin Histograms

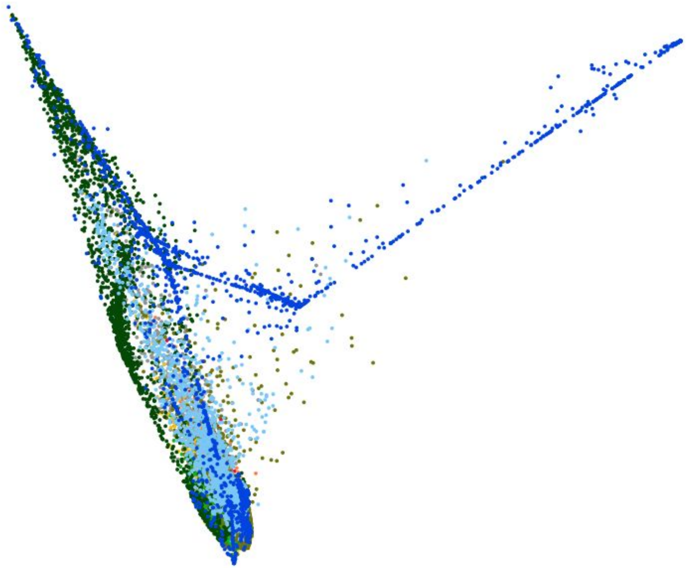
Original Image

K-Means Color Quantized - 64 Bins

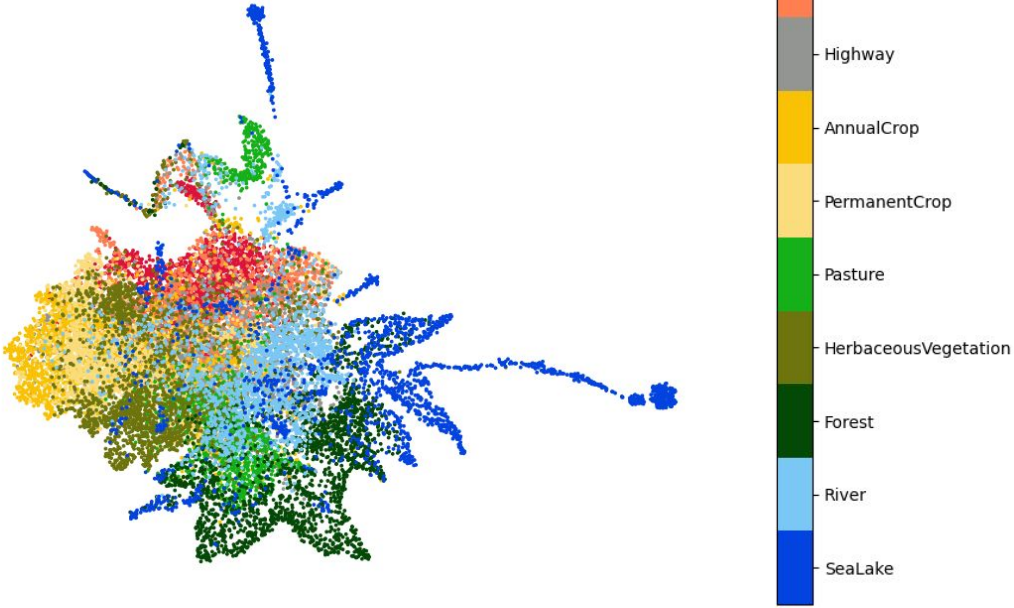


K-Means Color Bin Histograms

PCA

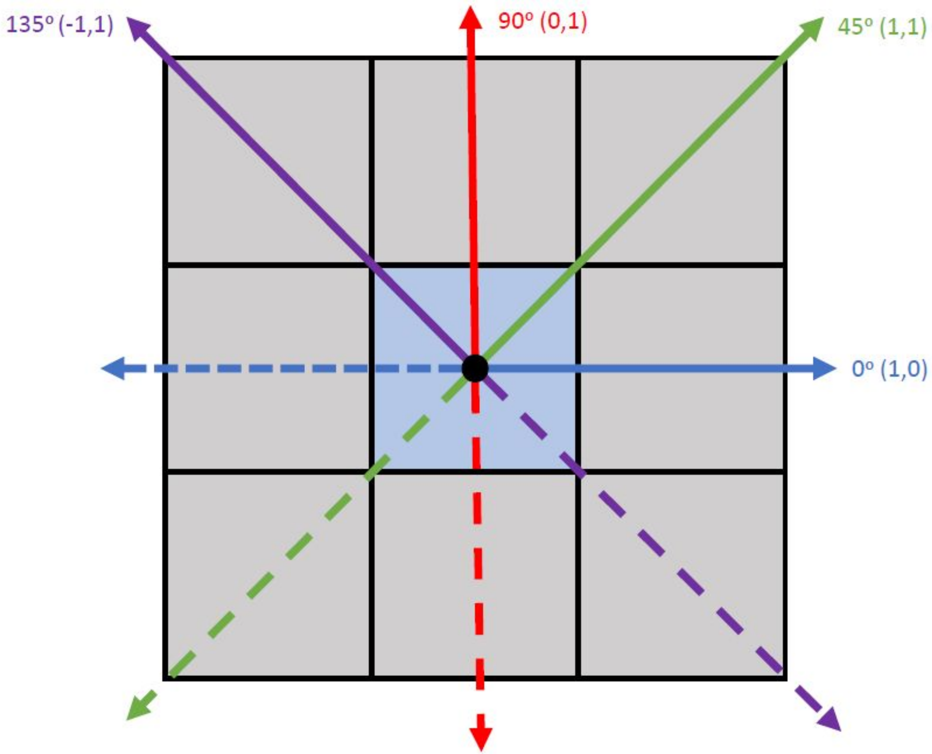


UMAP

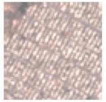


Gray-Level Co-occurrence Matrix (GLCM)

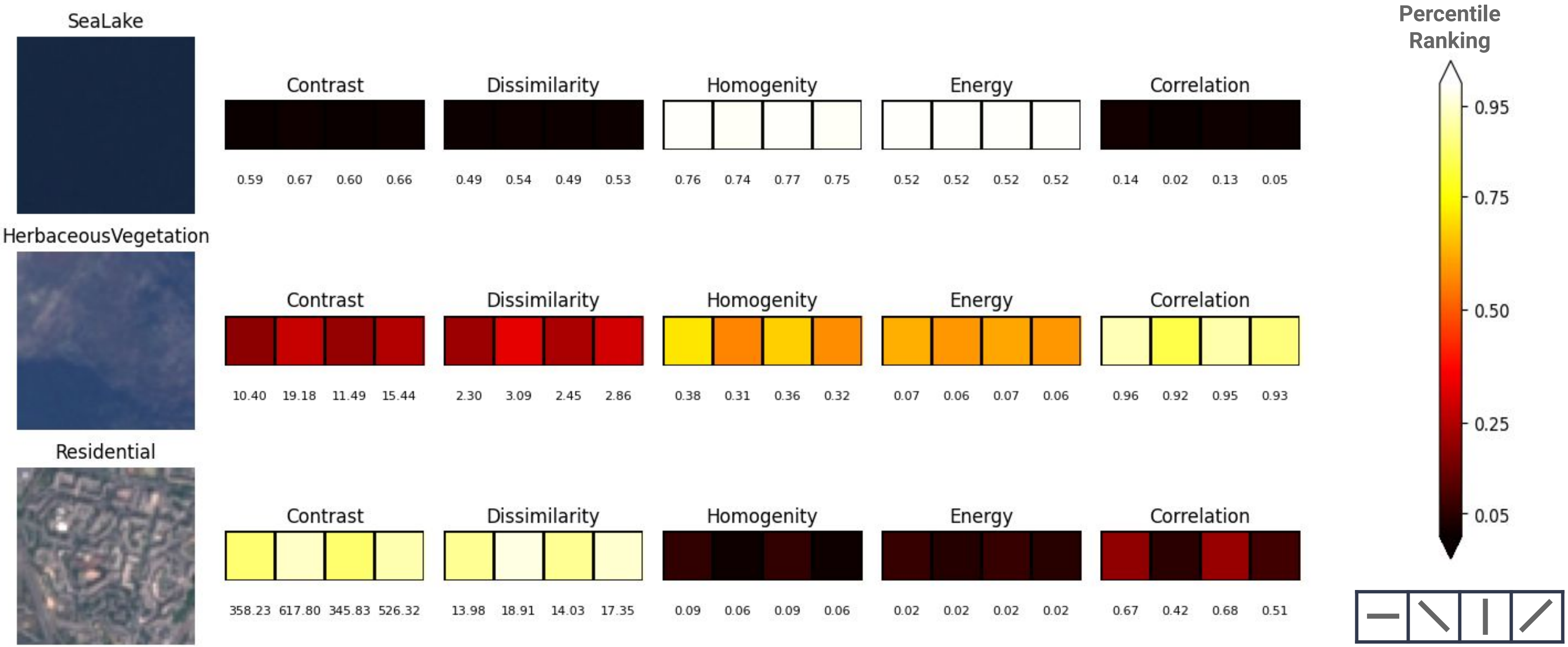
- GLCM is a statistical method of examining texture that considers the spatial relationship of pixels.
- The construction of GLCM needs two inputs: **pixel distance** and **direction**
- GLCM accounts for the frequency of pixel pairs occurring at the specified distance and direction.
- The matrix shape is determined by the number of gray levels, which in our case is (256, 256)
- Statistical properties of GLCM to include: **Contrast, Dissimilarity, Homogeneity, Energy, Correlation**



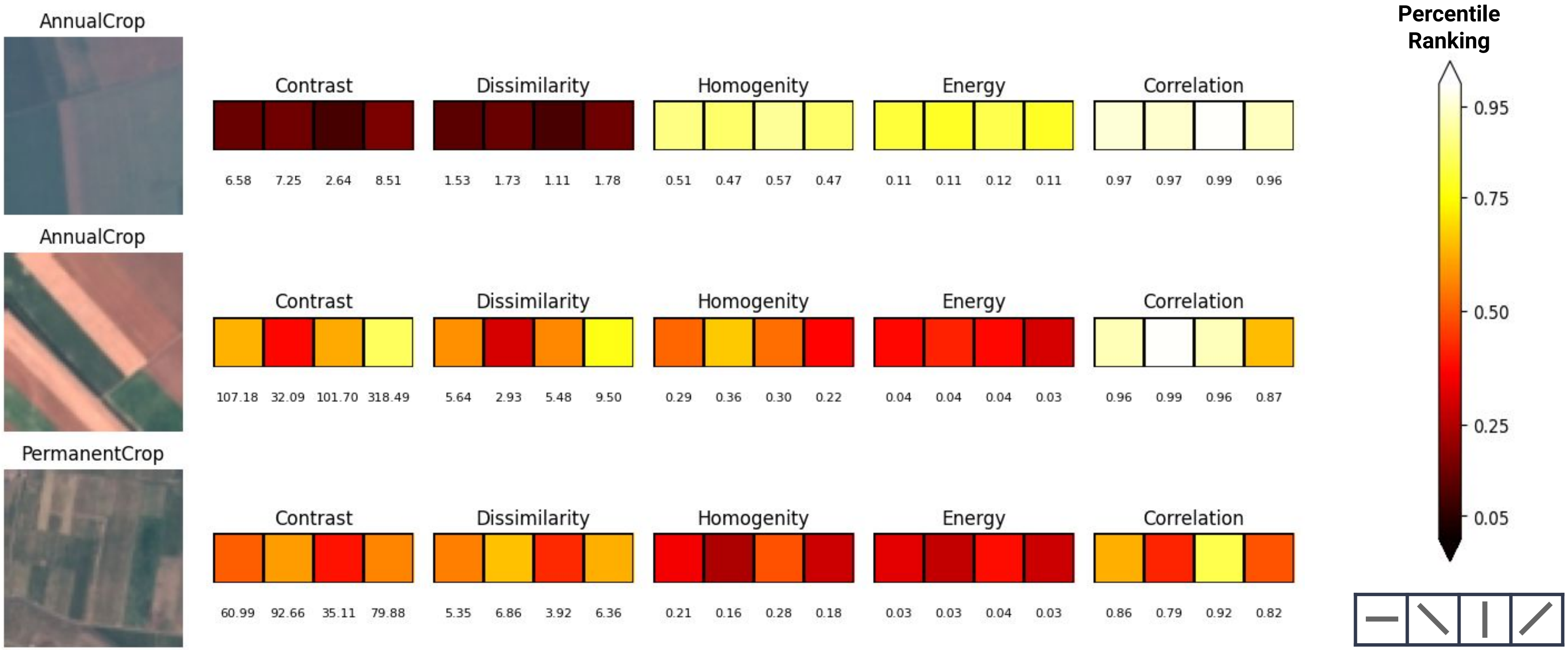
Gray-Level Co-occurrence Matrix (GLCM)

Feature	Description	Formula	Sample w/ Low Value	Sample w/ High Value
Contrast	Measures the local variations in the GLCM	$\sum_{i,j} P(i,j)(i-j)^2$		
Dissimilarity	Similar to Contrast, but takes the absolute difference, making it less sensitive to larger differences	$\sum_{i,j} P(i,j)\lvert i-j\rvert$		
Homogeneity	Measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal	$\sum_{i,j} P(i,j)/(1+(i-j)^2)$		
Energy (Angular Second Moment)	Sum of squared elements in the GLCM	$\sum_{i,j} P(i,j)^2$		
Correlation	Measures how correlated a pixel is to its neighbor over the whole image	$\sum_{i,j} P(i,j)((i-\mu_i)(j-\mu_j))/(\sigma_i\sigma_j)$		

Gray-Level Co-occurrence Matrix (GLCM)

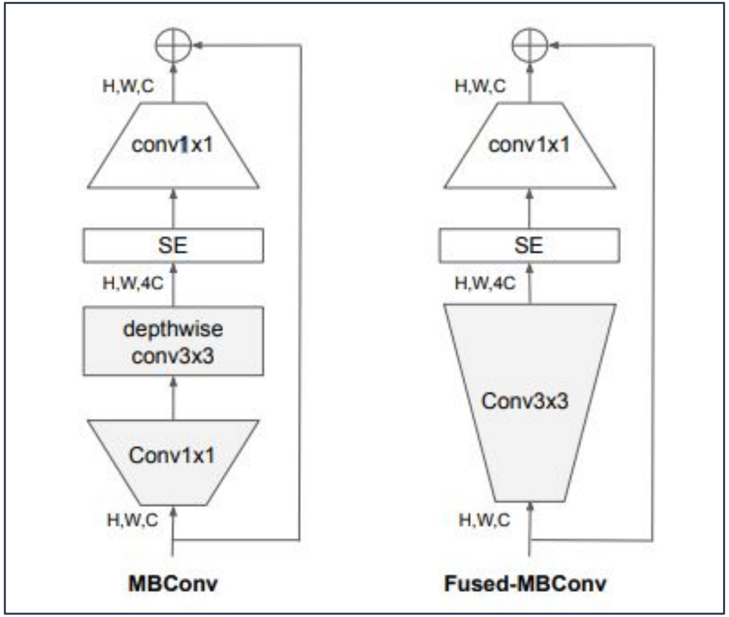


Gray-Level Co-occurrence Matrix (GLCM)

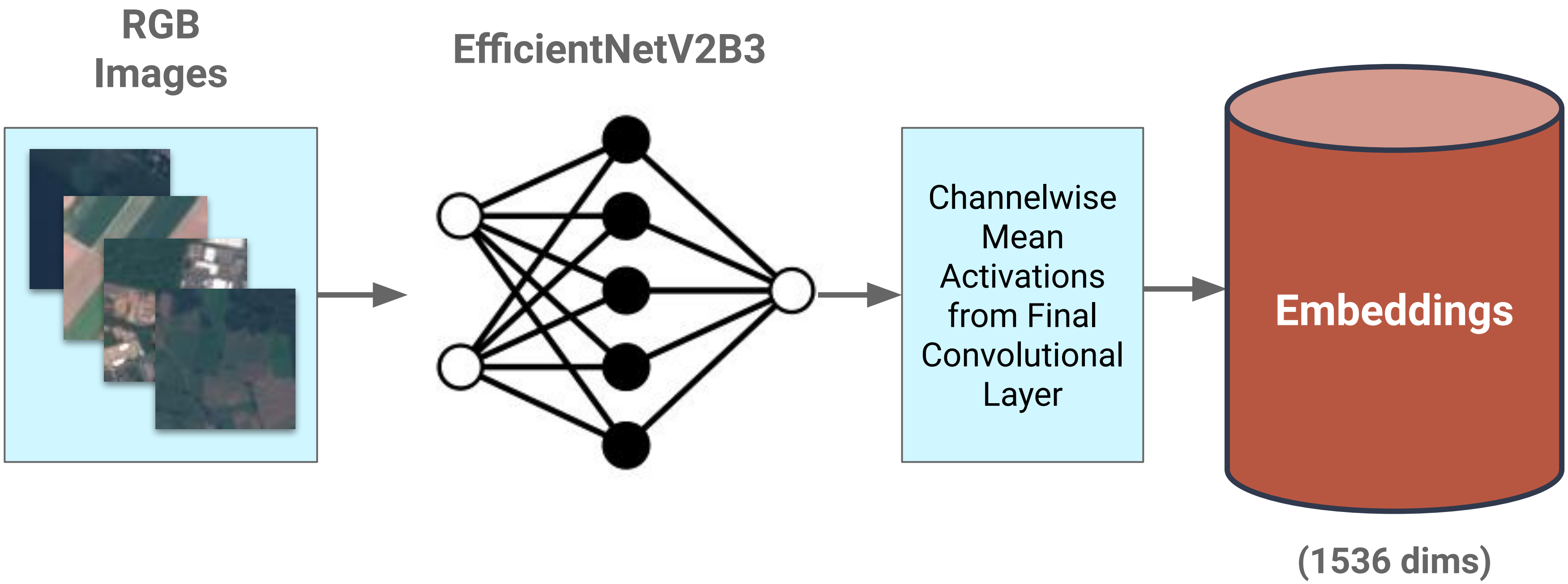


EfficientNet

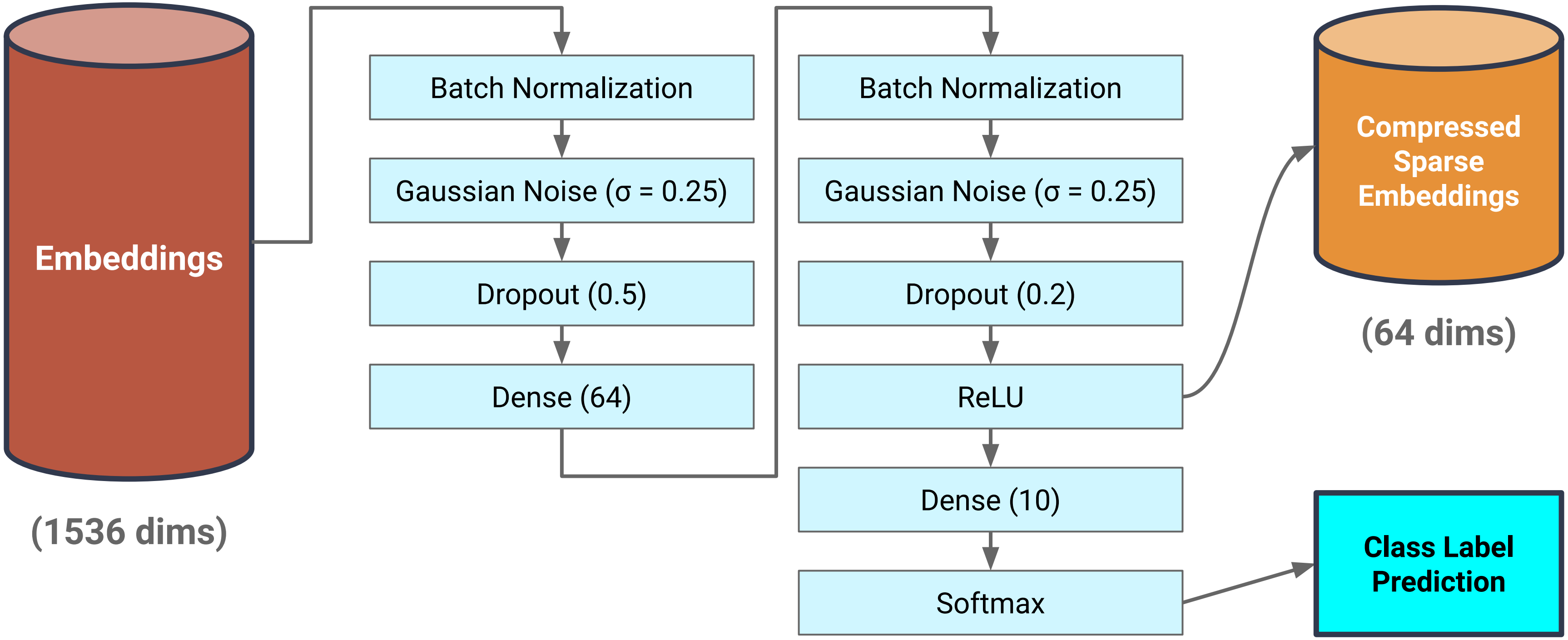
- CNN model architecture developed by two Google researchers
- Designed to be as accurate as SOTA image classifiers with fewer parameters, lower computational burden, and faster inference speed
- Final selection: **EfficientNetV2B3**
  - ImageNet-pretrained checkpoint available through Keras
  - Built-in pre-processing & transfer learning functionality
  - Can natively handle 64x64 resolution



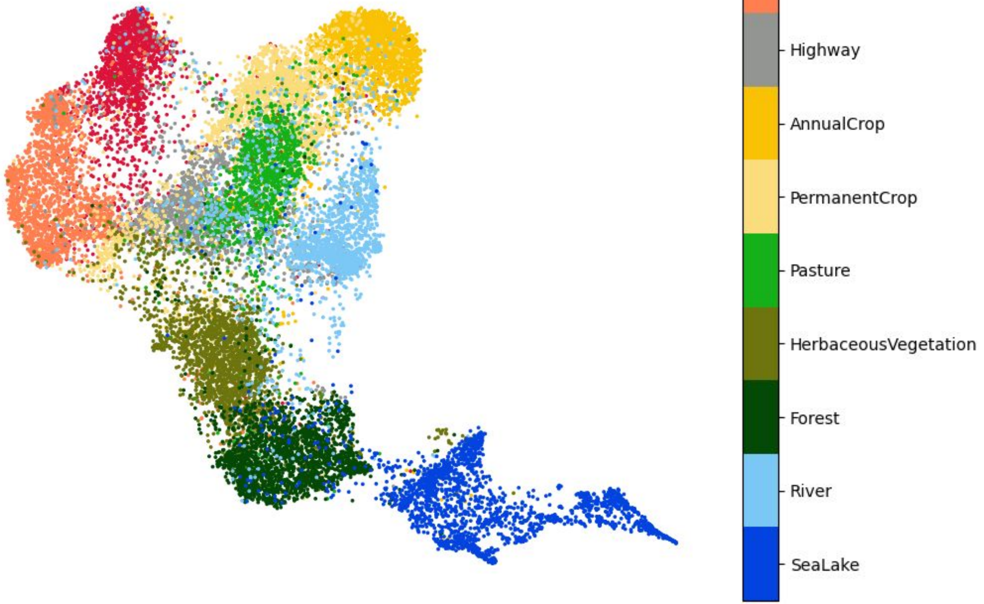
EfficientNet Embeddings



EfficientNet Embeddings – Compression

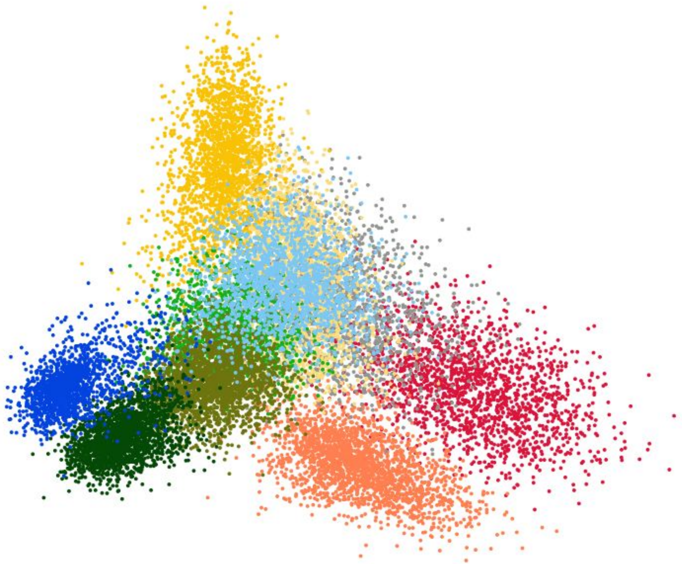


EfficientNet Embeddings

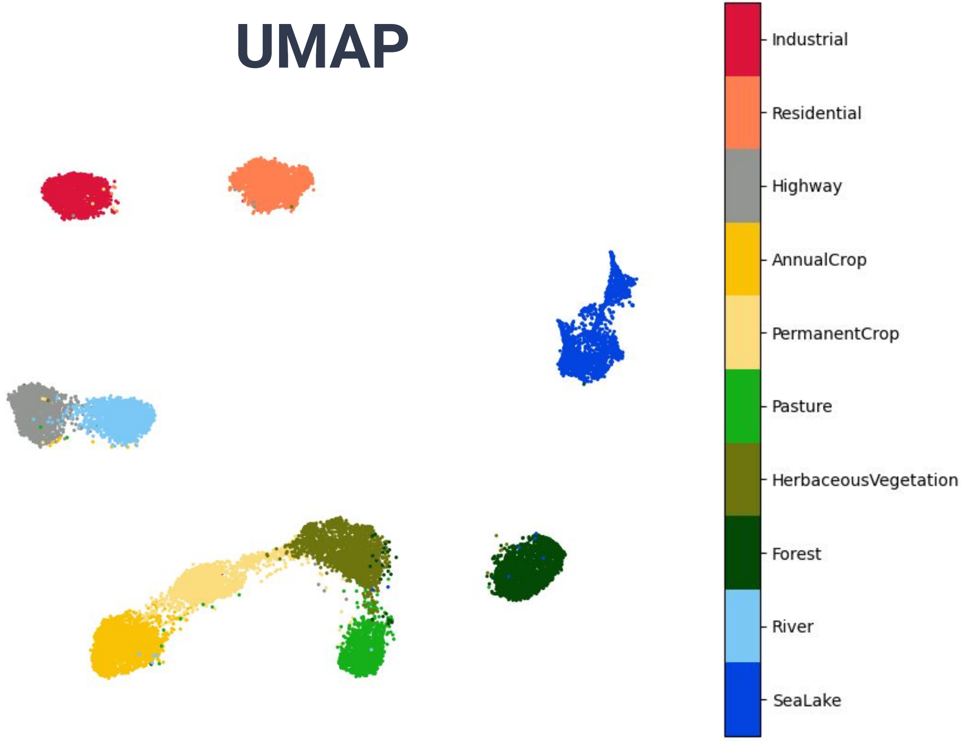


EfficientNet Embeddings - Compressed

PCA



UMAP



Model Architectures

Logistic Regression

- Multiclass classification using multinomial regression
- Pros:
    - Computationally cheap
    - Easy interpretability
    - Less prone to overfitting
- Cons:
    - Limited to linear relationships
    - Sensitive to feature scaling
- Searched **regularization coefficient (‘C’)** and **max iterations** parameters

eXtreme Gradient Boosting (XGBoost)

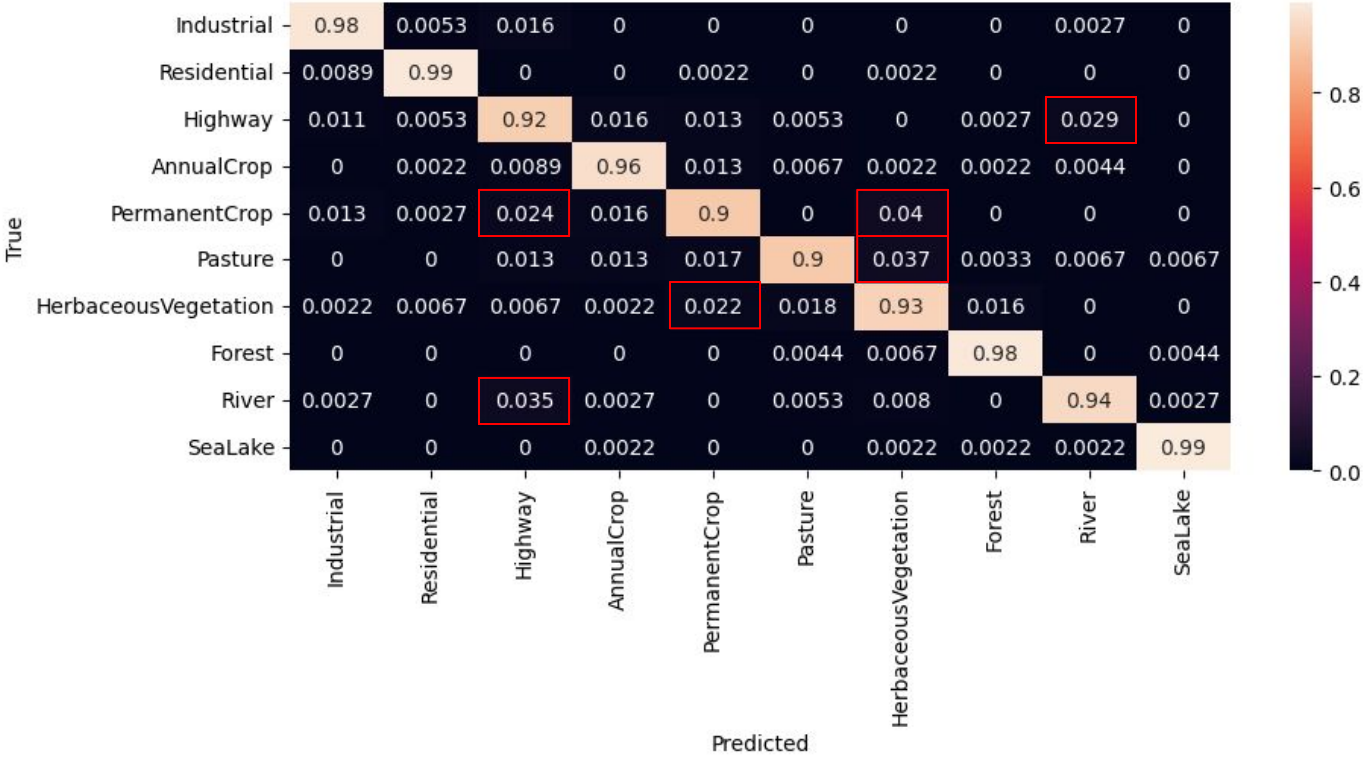
- Uses results of multiple decision trees to capture complex non-linear relationships between features and class labels
- Pros:
    - Incorporates gradient boosting and regularization
    - Early stopping
    - Handles non-linearity
- Cons:
    - Computationally expensive
    - Black box in nature
- Searched **n_estimators, max_depth, learning_rate** parameters

Preliminary Results

Model Type	Feature Set	Best Configuration	Training Time (sec)	Training Accuracy	Validation Accuracy	
Logistic Regression	Basic	C = 1000 max_iter = 3000	61.6	87.3%	85.7%	+ 9.5%
	Full	C = 1 max_iter = 100	2.36	99.3%	**95.2%**	
XGBoost	Basic	learning_rate = 0.1 max_depth = 7 n_estimators = 200	48.6	100.0%	92.8%	+ 1.9%
	Full	learning_rate = 0.1 max_depth = 3 n_estimators = 100	18.4	99.7%	94.7%	

Basic = Avg RGB Values, Kmeans Color Bin Histograms, GLCM Features = **107** columns
Full = Basic + EfficientNet features = **171** columns

Confusion Matrix – Test Set



PCA Analysis of Full Feature Vector