Landscape Classification

Kevin Dai, Gray Selby, Yencheng Chen

281 Final Project- Fall 2023

Github link

Abstract

In this report, we conduct the investigation of landscape images dataset from Kaggle:AID A Scene Classification Dataset. The dataset collected sample images from Google Earth around the world mainly in China, the United States, England, etc. Various features of the images are explored, from mean and variance of color channels (RGB, HSV), histogram of oriented gradients of gray, red, green, blue, hue and saturation channels to the pre-trained deep learning models ResNet and EfficientNet. Principal component analysis (PCA) was conducted to reduce the dimensionality of the feature sets into different experiments of feature combinations. With our ultimate feature set selected after a semi-exhaustive search, we achieved the accuracy of 0.914 in our final model.

1 Introduction

For our project, we classify the Google Earth imagery dataset, consisting of landscape images from the following 30 scene types: airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert, farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storage tanks and viaduct. Each class has 220 to 420 images, each image with dimensions of 600x600 pixel resolution. Fig. 1 shows some examples of the classes with a corresponding example image.

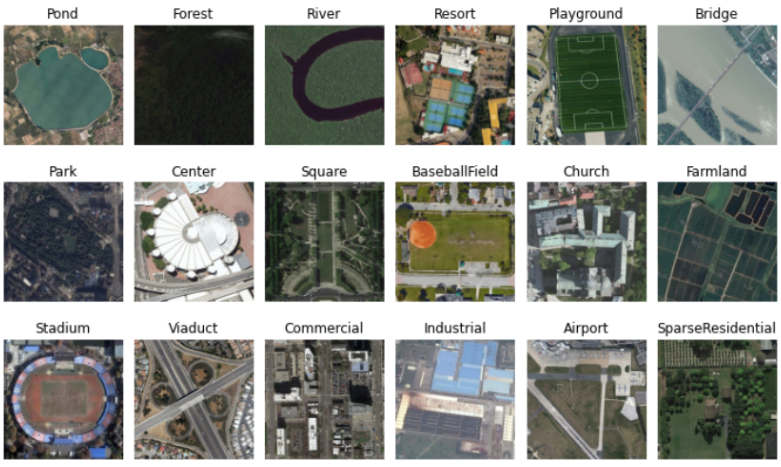


Fig. 1: Example images from the landscape dataset.

The images of this dataset are post-processed using RGB rendering from original optical aerial images. The images are multi-source, as Google Earth images are from different remote imaging sensors and the sample images are chosen from different countries and regions, different time and seasons and imaging conditions. This contributes to a significant amount of variance across classes, as well as within each class.

We consider low-level features such as RGB and HSV mean and variances and size-variant histogram of oriented gradients (HOG) for both grayscale and multichannel images. In addition, we consider higher level features derived from the weights of pre-trained deep learning models such as VGG and ResNet. To optimize our final feature set, we perform experiments with various combinations of these features using cross-validation. In order to reduce dimensionality and improve efficiency while attempting to preserve explained variance, we also experiment with principal component analysis (PCA) in conjunction with our combinatorial feature search.

2 Feature Extraction:

The feature extraction for the classification is explored. The motivation and effectiveness of different choices are discussed below:

2.1 Mean and Variance of RGB Channels

From the general observation, classes of scenes included different types of color tint in the image. For example, the classes Forest and Mountain have a majority of green pixels, while the Port and Beach classes have majority blue. For a low level feature, we believe that the mean value of the red, green, and blue channels in each image can provide initial predictive power on its classification, but may not be able to differentiate between images with similar color profiles, such as the aforementioned Forest and Mountain classes. Thus, we also introduce the variance of each channel to assist with capturing this variability. This differentiating power can be seen in Fig 2, where the first three plots show the distribution of mean values for the RGB channels and the bottom three show the variance. While there are some classes that show distinct distributions for mean RGB values, it is clear that variance can help with additional separation. However, the distributions of the three channels with each class do not seem to be particularly varied; that is, a class with high variance for one channel likely has high variance for the other

two as well, and same for the mean values. This may not be ideal for modeling as it is possible that there is some covariance introduced by considering all of these variables together.

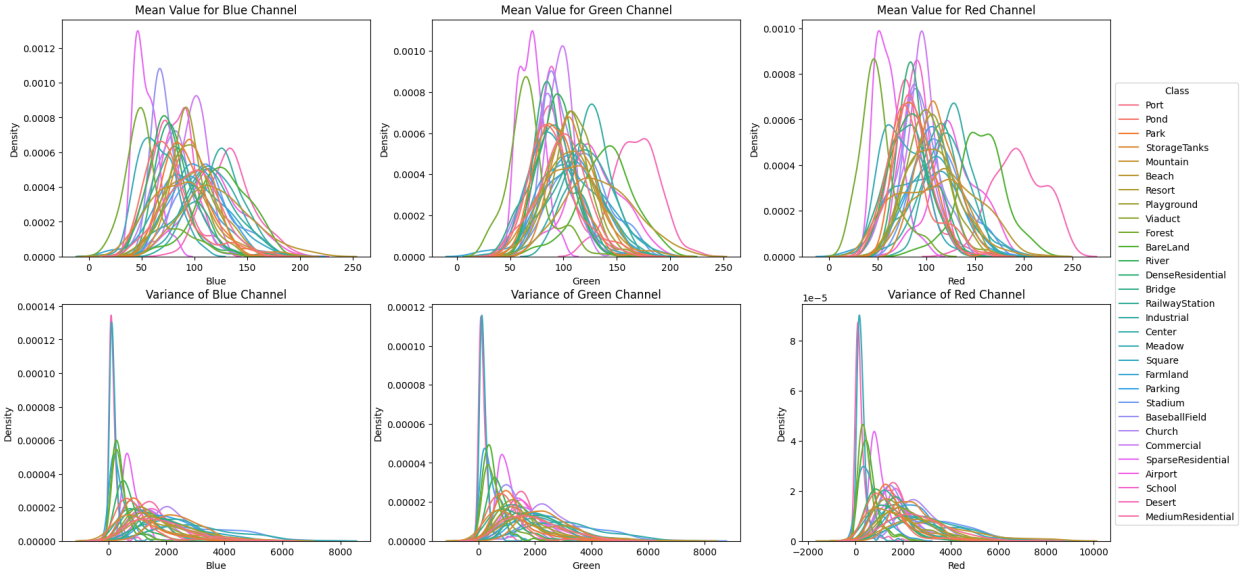


Fig. 2: Pixel RGB distributions. The top row shows distributions for mean values for each channel and the bottom row shows distributions for variance.

2.2 Mean and Variance of HSV Channels

In addition to RGB values, we also consider similar features for the hue-saturation-value (HSV) color model. Because landscape images may contain a multitude of shadows and other terrains, the HSV color model will be a cleaner representation due to the color invariance on the hue channel. For many of our classes, the ability to differentiate based on luminosity and saturation, as well as keeping hue on a single continuous scale will prove to be valuable. This can be seen in our initial exploration of the distributions of HSV mean and variance within each class in Fig 3. While it is still difficult to completely differentiate between each class, the hue and brightness channels in particular have somewhat more distinct class distributions as compared to the RGB features in Fig 2.

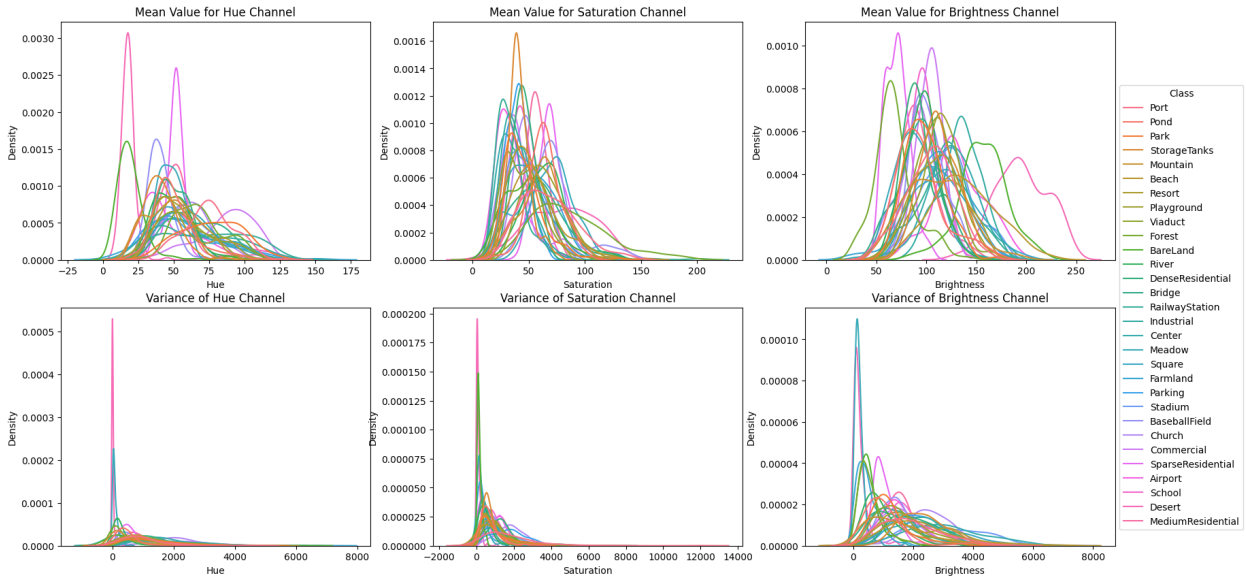


Fig. 3: Pixel HSV distributions. The top row shows distributions for mean values for the hue, saturation, and brightness channels and the bottom row shows distributions for the respective variance.

2.3 Histogram of Oriented Gradients (HOG):

We consider histogram of oriented gradients (HOG) in an attempt to capture the significant objects within our images. For many of our classes, there are defining objects with edges such as winding roads, airplanes, stadiums, or houses. Due to HOG's ability to not only identify the edges of objects but also the edge directions and magnitudes, it is a great candidate for extracting differentiating features. In order to extract the edges of objects in our images, we tried different sizes of pixel cells and numbers of orientation to capture large and small scales of features. This is because the features vary in pixel scales; in some cases such as the differentiation of the Forest and Meadow classes, small pixel cells could be more effective. On the other hand, separating classes such as School and Port would be more effective with large pixel cells. Also, different landscapes have various gradients in terms of color or HSV channels which might be useful to separate classes like the Forest and Meadow. To accommodate this, we not only create HOG vectors for the grayscale image, but also for each individual RGB and HSV channel. This will potentially allow the classification model to differentiate between classes with similar color schemes by being more granular when we capture features. In our model we picked two kinds of HOG size, small: orientation = 8, pixels_percell = (20x20) and large: orientation = 8, pixels_per_cell = (50,50).

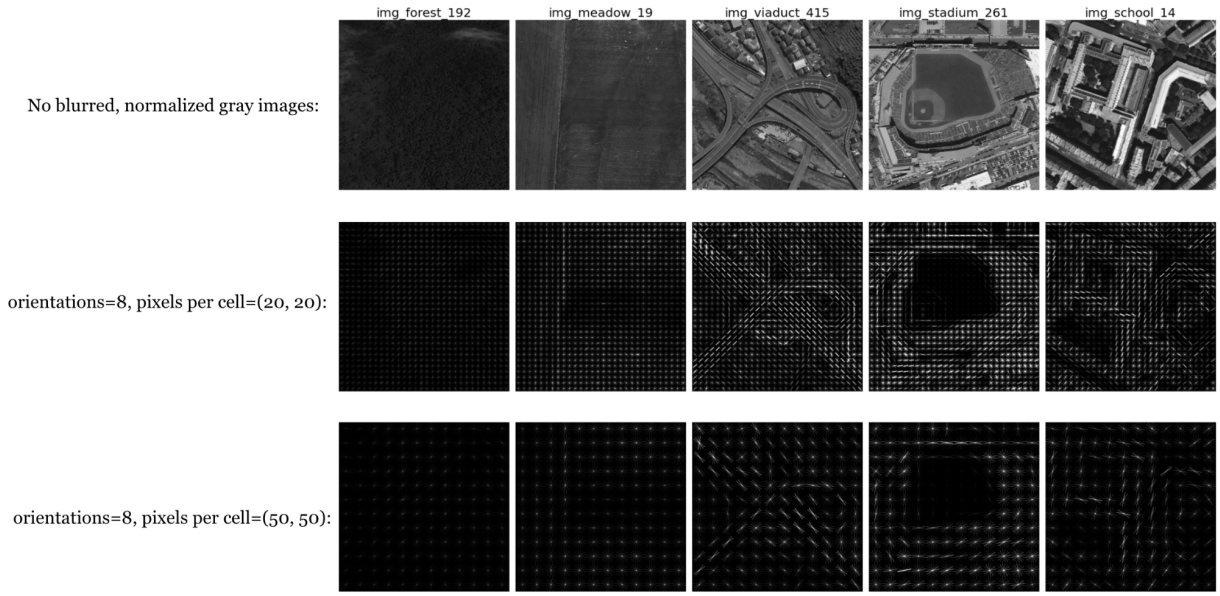


Fig. 3: Gray scale HOG with different cell parameters

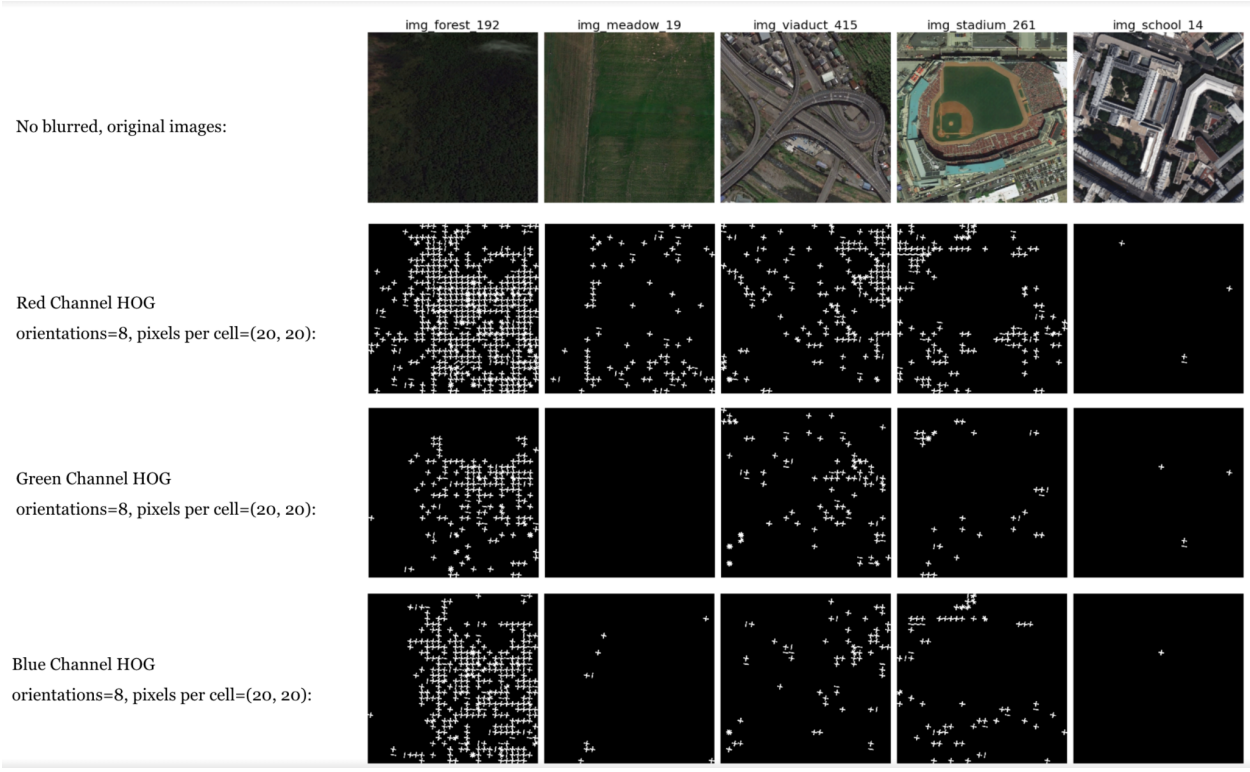


Fig. 4: RGB HOG with different cell parameters

2.4 DAISY Descriptor

The DAISY descriptor, similar to SIFT, is used to find the invariant for scale change or rotation. As a descriptor, it converts local image regions into low dimensional invariant descriptors which can be used for matching and classification. It can be applied densely or applied to patches extracted around interest points. The images in our dataset are all shot at the same angle(top) with different scales and resolution, meaning that DAISY can help provide useful features if the images are zoomed in or zoomed out. In our model, we picked step = 220, radius = 70, rings = 3, histograms = 8 and orientations = 8 to generate features. The visualizations of some of these features can be seen in Fig 5

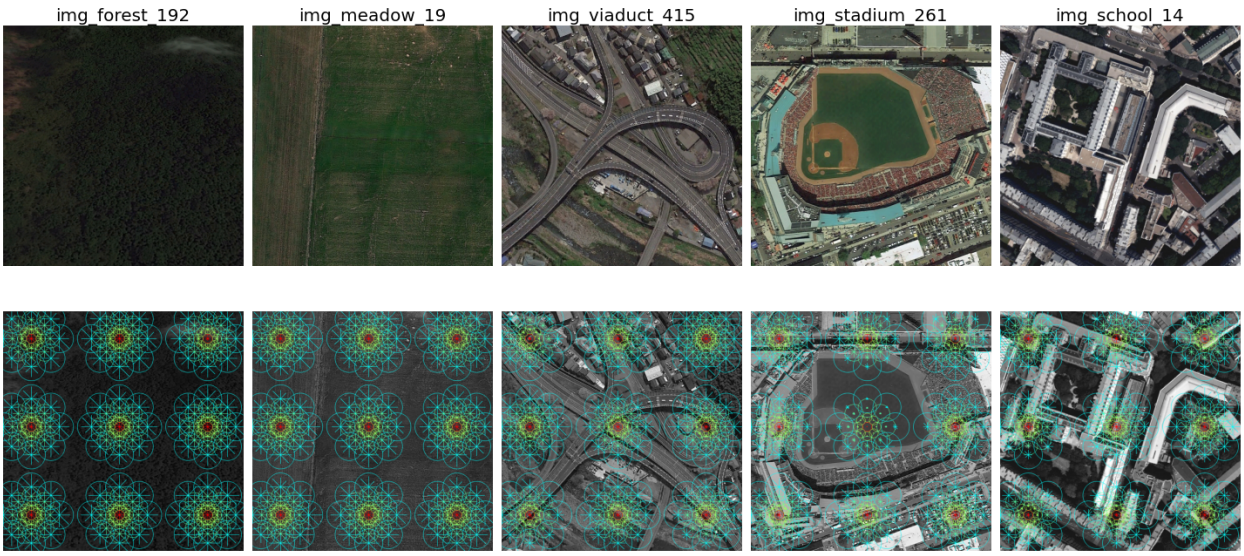


Fig 5: Extracted DAISY features

2.5 Pre-trained Neural Network Model ResNet Feature Embeddings:

For high-level complex features, we consider embeddings from two pre-trained deep neural networks as feature extractors: ResNet101 and EfficientNet, both of which are convolutional neural networks.

2.5.1 ResNet101

Residual Network-101 (ResNet101) is a CNN that has reformulated layers for learning residual functions with reference to the inputs to those layers. This allows for easier optimization and results in significantly increased accuracy the deeper the network is. For this version of the model, ResNet101 is trained with 101 layers on ImageNet-1k at a resolution of 224x224 pixels and it is slightly more accurate than its predecessor at the cost of a small efficiency drawback.

For our use case, we scale down our images to 224x224 and use them to extract the embeddings from ResNet101 right before the classification layer, which we can use as standalone features for our model and in combination with our lower-level features to gain improved performance.

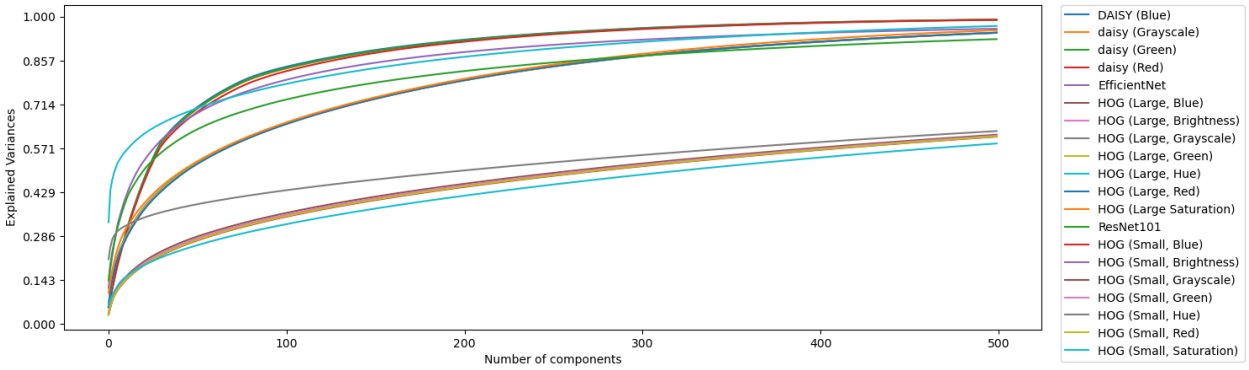
2.5.2 EfficientNet (size medium)

Similar to ResNet101, EfficientNet is a CNN that attempts to achieve high accuracy in image classification tasks. However, EfficientNet aims to achieve top performance while preventing the hyper-scaling of layers, as is commonly accepted as a good and simple way to improve performance. By utilizing compound coefficients, EfficientNet uniformly scales each dimension with a fixed set of coefficients and is able to achieve state of the art performance while being over eight times smaller and six times faster than other models.

We similarly rescale and extract the final embedding layer of the pre-trained medium sized EfficientNet model, which we are able to utilize individually or concatenate with our lower-level features.

2.6 Principal Component Analysis (PCA)

With our low-level and high-level features taken into consideration, our feature set is extraordinarily large, even with just a few combinatory features. Thus, we turn to principal component analysis (PCA) to reduce our dimensionality. PCA is able to accomplish this by transforming the dataset by projecting it into a new coordinate system while preserving the maximum amount of variance. One drawback of PCA is the effectiveness is highly dependent on the data itself. If the data is sufficiently noisy, not enough variance will be captured in a few number of principal components, significantly decreasing the efficiency. We evaluate the effectiveness of PCA on each of our features as well as our feature combinations by examining the explained variance plots shown in Fig 6.



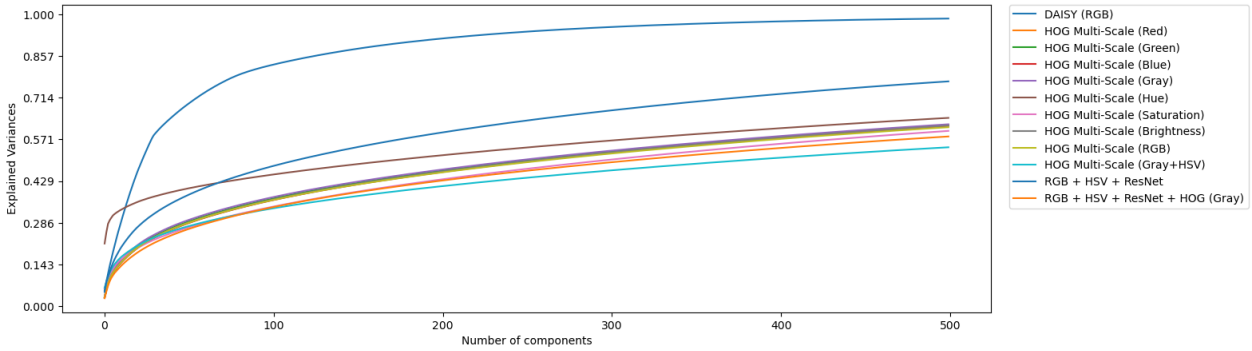


Fig 6: Explained variance of principal components of our large considered features

2.7 T-distributed Stochastic Neighbor Embedding (t-SNE)

T-distributed stochastic neighbor embedding (t-SNE) is another dimensionality-reduction method that projects data into a low-dimensional space in a way that similar points are near each other and dissimilar points are distant with high probability, determined by minimizing Kullback-Leibler divergence. Unlike PCA which prioritizes the optimization of the number of components and the explained variance, t-SNE is used to project the data into two or three dimensions only in order to visualize the data in a lower dimensional space. Fig 7 shows t-SNE plots for our various considered features. It is clear that ResNet has the cleanest clustering, followed by the combination of RGB and HSV features, which is to be expected as the learned embeddings are very robust and can accommodate the intra-class variance of our dataset.

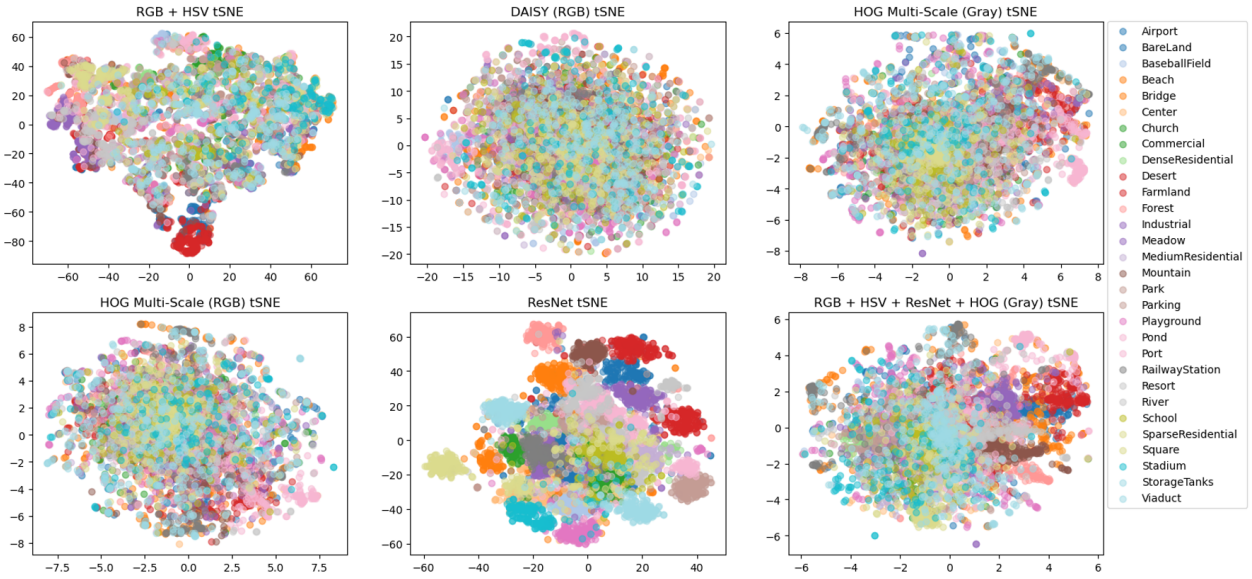


Fig 7: t-SNE visualizations of select features

3 Classification

Prior to developing our final classification model, we perform a feature selection process to select a subset of features with the highest predictive power. To do this, we perform a semi-exhaustive search of our features individually and in combination by training basic, unoptimized logistic regression models with 30-fold cross-validation. A subset of the results can be found in Table 1.

Feature Name	Accuracy	F1-score	Train time (seconds)	Evaluation time (seconds)
RGB + HSV	0.42	0.40	0.94	0.00
DAISY (RGB)	0.27	0.27	143	0.03
HOG Multi-scale Grayscale	0.32	0.31	245.92	0.05
ResNet	0.92	0.92	54.14	0.02
EfficientNet	0.91	0.91	33.2	0.01
RGB + HSV + ResNet	0.91	0.91	54.2	0.01
PCA(0.95): RGB + HSV + ResNet	**0.91**	**0.91**	**7.88**	**0.01**
RGB + HSV + ResNet + HOG (Grayscale)	0.87	0.87	297.5	0.05
PCA(0.95): RGB + HSV + ResNet + HOG (Grayscale)	0.87	0.87	18.5	0.01
…	…	…	…	…

Table 1: Feature selection results. A subset of a representative portion of our feature combinatory set is shown here, out of 44 total experiments

Ultimately, we decided to use a combination of RGB, HSV, and ResNet features with PCA applied on top, taking into account the high performance as well as the efficiency of the training and evaluation.

3.1 Results

We finally train 2 models: 1) Logistic Regression, and 2) Support Vector Machine (SVM). Logistic regression is a probabilistic classification model that outputs class probabilities using a sigmoid, whereas SVMs allow for a more spatial approach, creating boundaries between classes in

high-dimensional space by maximizing the margins for each class. We apply both methods and evaluate the best model. For each model type, we perform hyper-parameter optimization through gridsearch and 5-fold cross validation. Our results can be found in Table 2.

Classifier	Training Accuracy	Validation Accuracy	Test Accuracy
Logistic Regression	1.0	0.91	0.914
SVM	1.0	0.90	0.913

Table 2: Final model results.

3.2 Selected Model

We select Logistic Regression as our final model, as it displays a slight performance improvement over the SVM on the test set performance. Our confusion matrix for the models can be found in Fig 8.

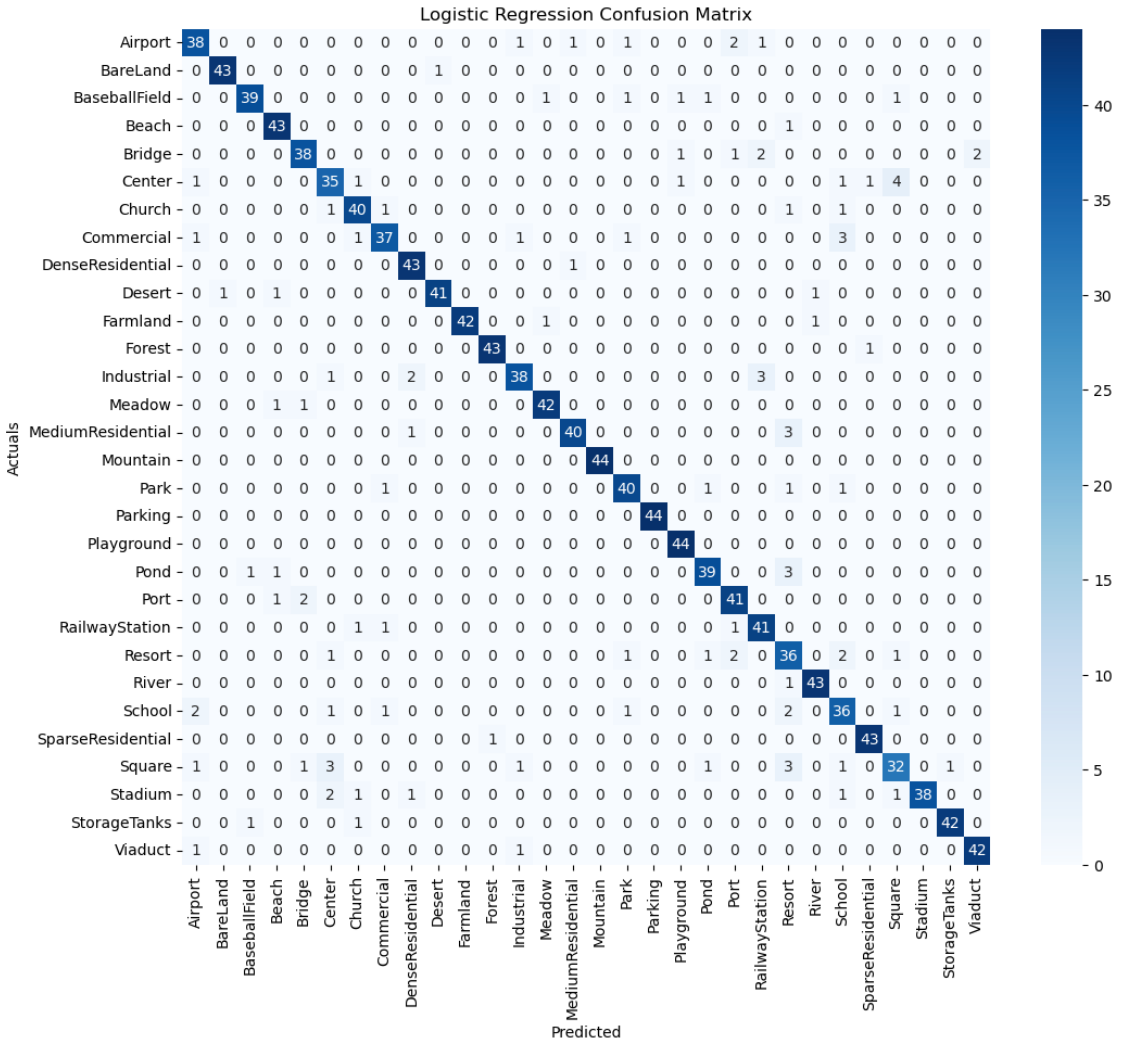


Fig 8: Confusion matrix for our Logistic Regression (final) model.

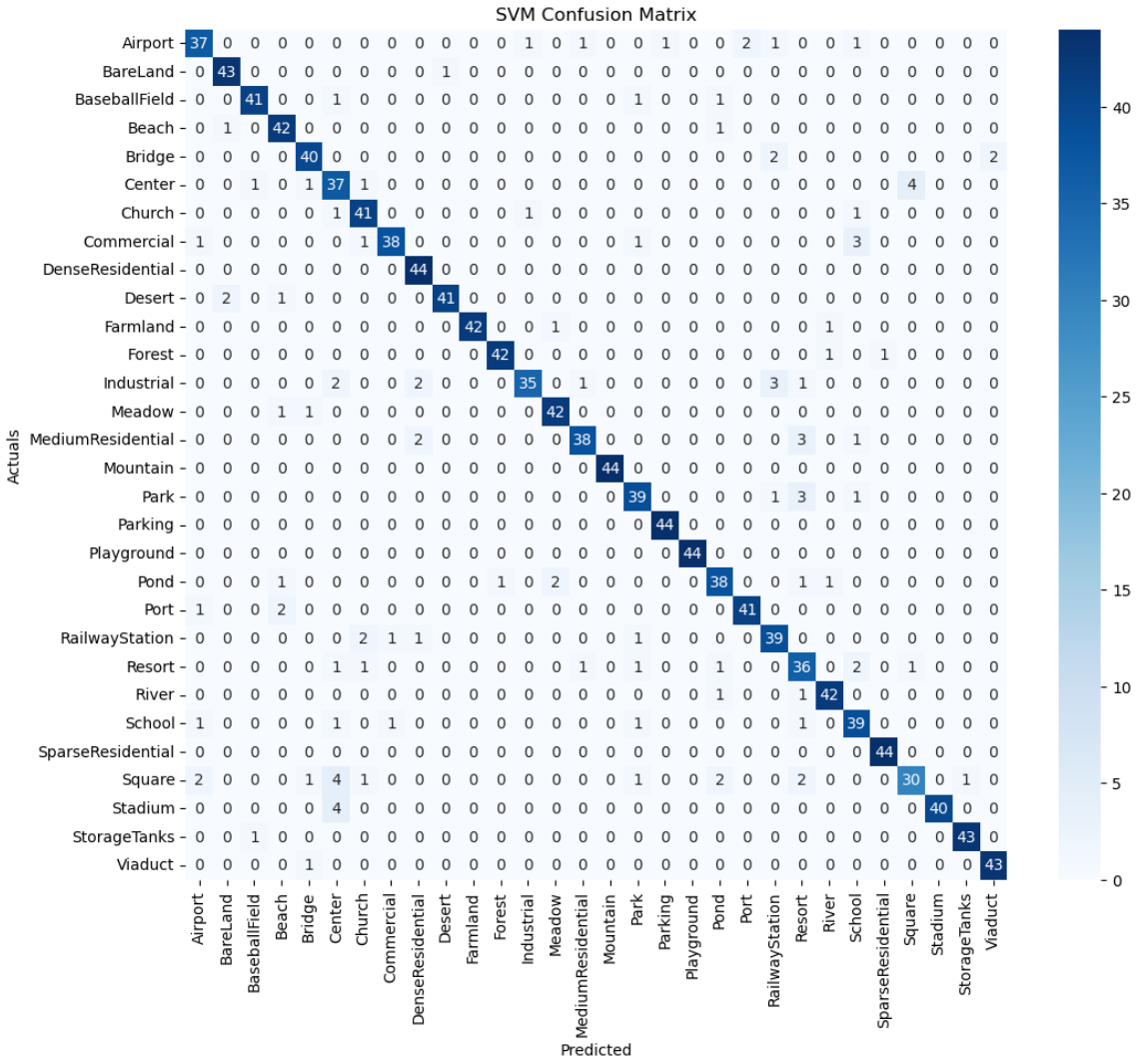


Fig 9: Confusion matrix for our SVM model.

3.3 Example of Misclassifications

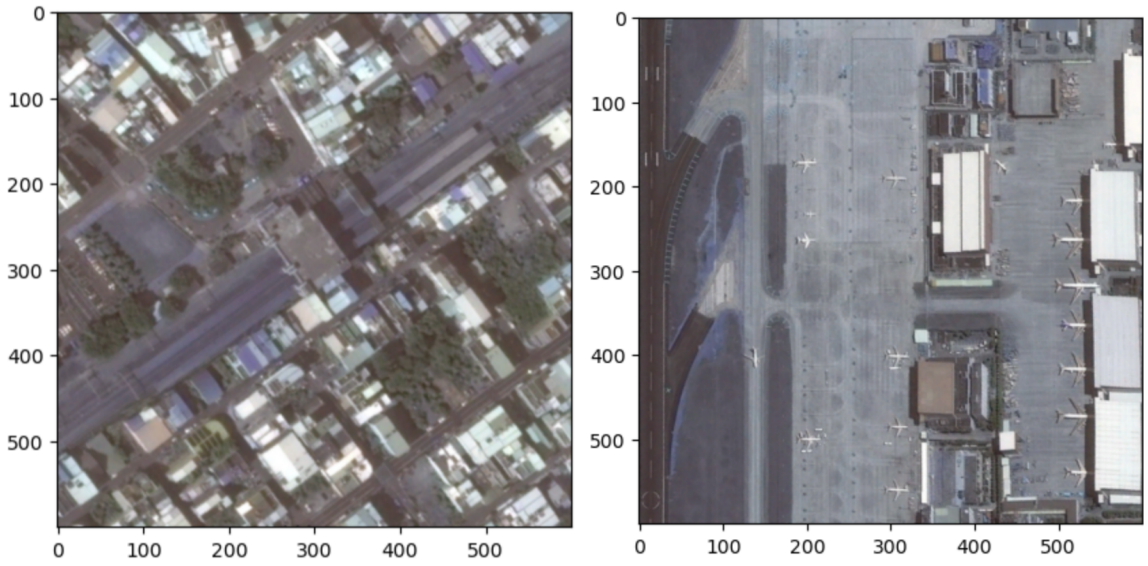


Fig. 10

Fig 10-1: Class RailwayStation, but it was pred as Commercial
Fig 10-2: Class Airport, but predicted as Industrial

We can see from investigating misclassifications that there are many similar classes. In these two examples, we can see that it might be hard for a human to get the first misclassification correct as the RailwayStation looks like it could be Commercial. On the other hand, a close look at the second misclassification reveals small airplanes that would help to classify this correctly as an Airport. This suggests that even more fine grained HOG or DAISY features could improve on some of these misclassifications.

4 Generalizability

Following common practices, the dataset was split into 80%, 20% for train and test datasets, and experiments on the train dataset were conducted and validated by multi-fold cross validation and confusion matrices to make our decisions of features selection. The test dataset was kept intact until the final feature vector and classification. Although overfitting is a concern, evaluation on the test set shows that the selected models generalize well to the test set.

5 Efficiency vs. Accuracy

To understand the efficiency of the classifiers, the training and evaluating time are recorded across all models. In the combined features with PCA applied, the validation is a bit worse, although that is expected since our reduction in the number of features will result in some explained variability

lost. However, this reduction in dimensionality has significantly improved efficiency, as can be seen during our feature selection process. We also theorize that the PCA features will result in less overfitting, but that will require future work to further quantify this.

6 Conclusion

From our exploration and experimentation, low level features such as RGB, HSV, HOG with different scales and DAISY combined may not provide us high accuracy as we expected. For example, some images in the Airport class have few to no airplanes within tiny portions of the images, it would be easily messed up by roads or buildings. To address this problem, I think we should try another approach of combining some datasets which are naturally hard to separate. For example, in some pictures within the SparceResidential class, there are mostly green areas in between buildings, which could potentially be misidentified as Park. In addition, BareLand is no more than dirt and open space in the entirety of each image, which is very similar to Desert. Similarly, School and Commercial images are both dense clustering of several large buildings. In terms of time time limitation and computation capability, we will need more time to try out if this approach could provide better results. On the other hand, with the help of neural networks features in ResNet, we can achieve the accuracy to 91.4% and 91.3% of Logistic Regression and SVM separately in the test dataset.

References

[1] Kaggle dataset:
https://www.kaggle.com/datasets/jiayuanchengala/aid-scene-classification-datasets
[2] HOG feature:
https://www.analyticsvidhya.com/blog/2019/09/feature-engineering-images-introduction-hog-feature-descriptor/
[3] DAISY features: https://www.epfl.ch/labs/cvlab/software/descriptors-and-keypoints/daisy/