**University of Science and Technology of Hanoi**

The Vision Transformer for remote sensing image classification.

Group members:

Nguyễn Thành Nam - 22BI13045

Trần Trọng Nghĩa - 22BI13332

Trần Đức Dương - 22BI13117

Nguyễn Tuấn Dũng - 22BI13107

Trương Đại An - 22BI13008

Vũ Hùng Anh - 22BI13045

**September 2024**

**Introduction**

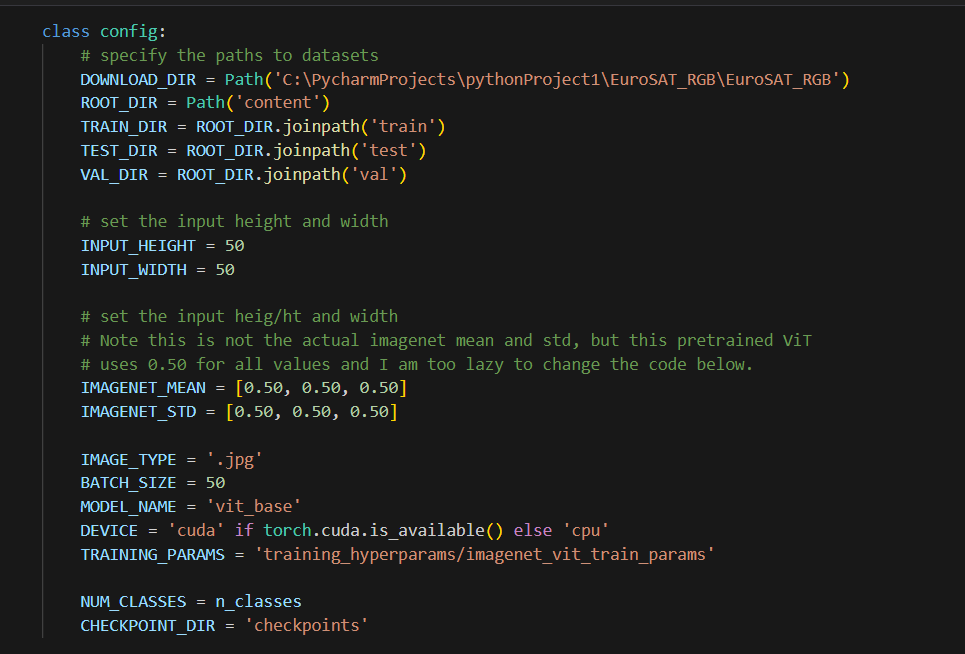
This report examines the application of Vision Transformers (ViT) for satellite image classification. ViT models leverage the transformer architecture, offering an innovative approach to image processing by dividing images into patches and treating them similarly to sequences in NLP tasks.

**Data Overview**

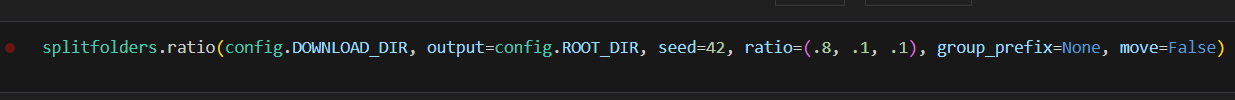
The dataset consists of high-resolution satellite images, each labeled with specific land cover types such as urban areas, forests, or water bodies. The data is divided into training, validation, and test sets to ensure robust model evaluation.

*Dataset Characteristics*

* Number of Classes: Multiple categories representing different land cover types.
* Image Dimensions: Varies, resized during preprocessing.



* Split Ratio: Typically, 80% training, 10% validation, 10% testing.



**Exploratory Data Analysis (EDA)**

*Plotted image*

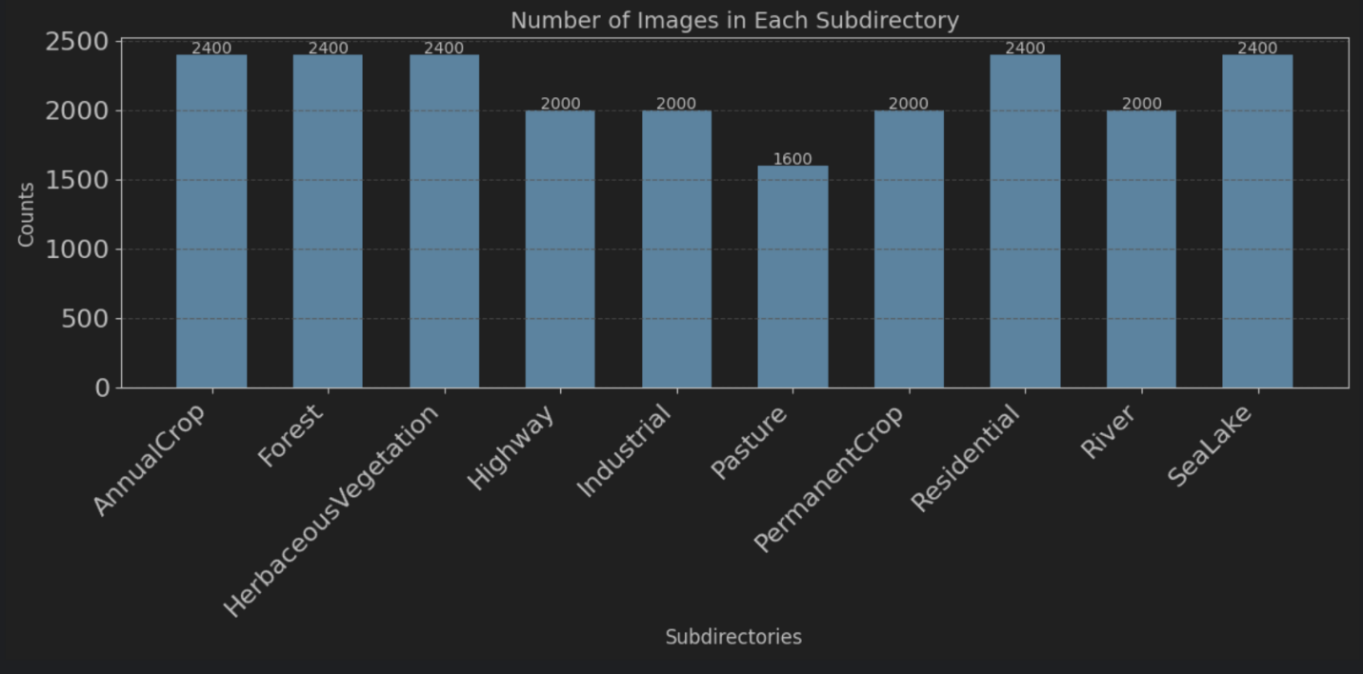
This helps visualize a few examples of the dataset's images to gain insight into their structure and content:



*Class distribution*

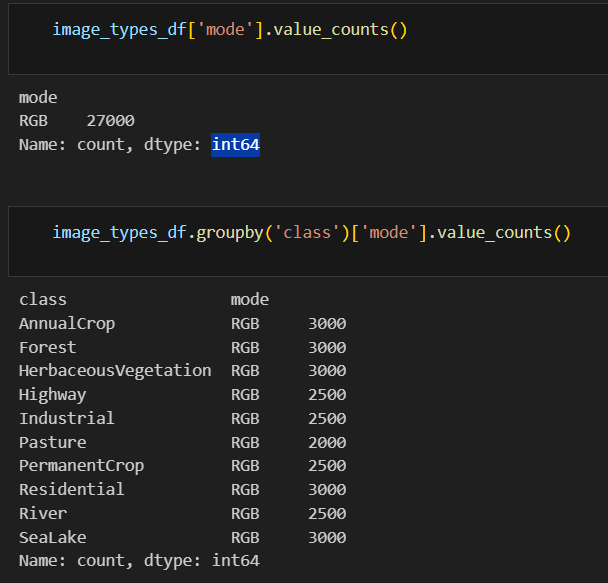
This involves analyzing how evenly the different classes (labels) are distributed across the dataset, which is crucial for detecting any imbalances.

*Number of image distribution*



*Image types*

Ensures the images are consistent in format and resolution also count and group them:



**Methodology**

*Vision Transformers (ViT)*

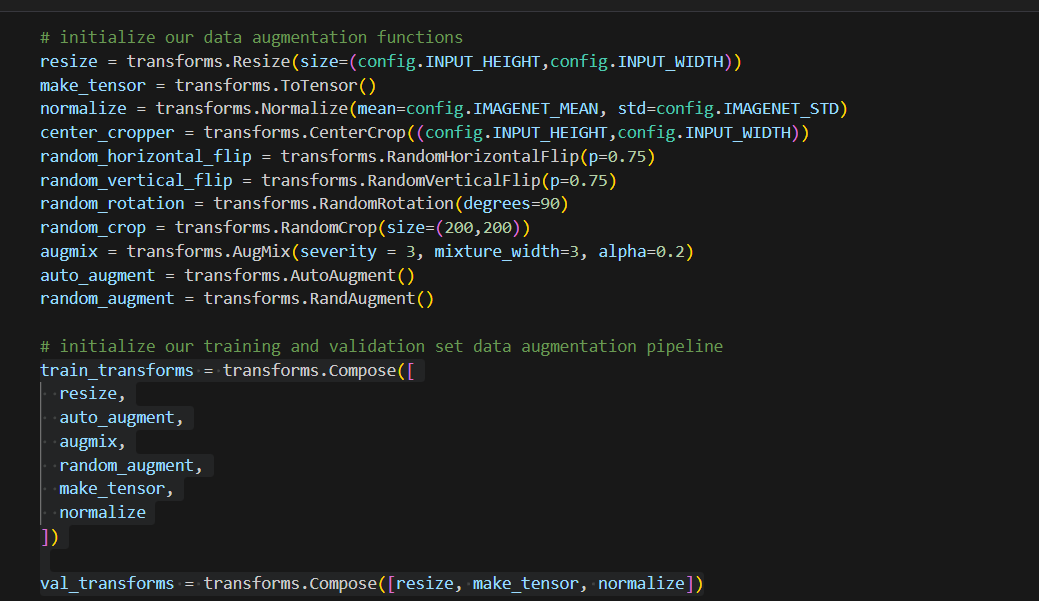
ViT models process images by splitting them into fixed-size patches. Each patch is linearly embedded, and positional embeddings are added to retain spatial information. These embeddings are then passed through a transformer encoder.

*Key Components*

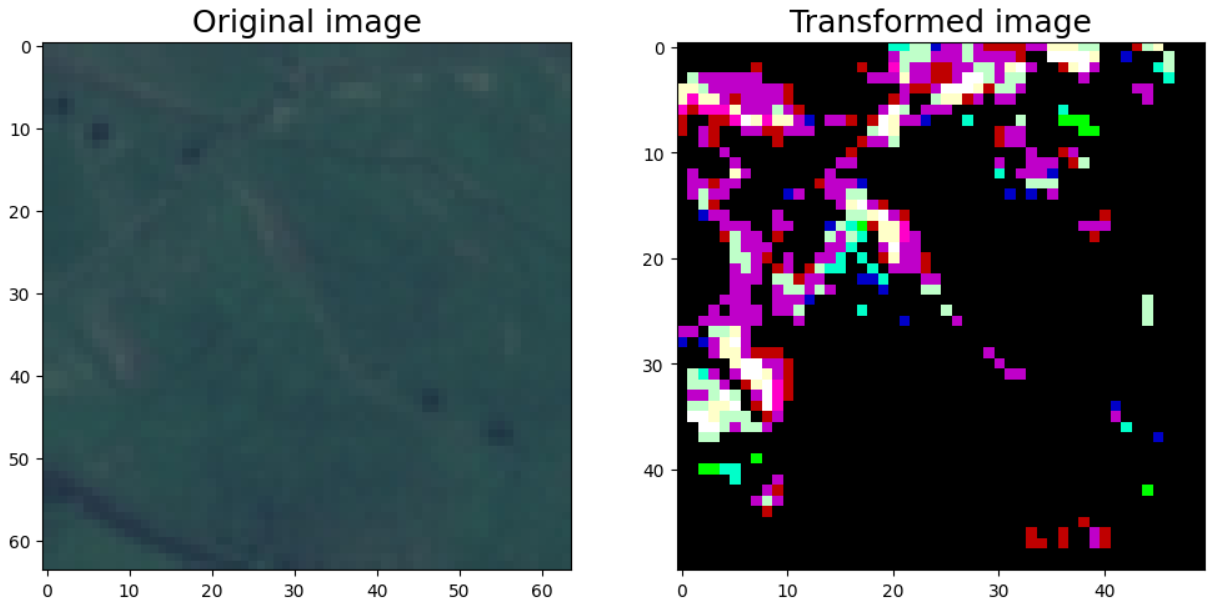
* Patch Embedding: Converts image patches into a sequence of embeddings.
* Transformer Encoder: Consists of multiple self-attention layers and feed-forward networks.
* Classification Head: Outputs class probabilities.

*Preprocessing*

Images undergo preprocessing to standardize input size and enhance generalization through data augmentation.



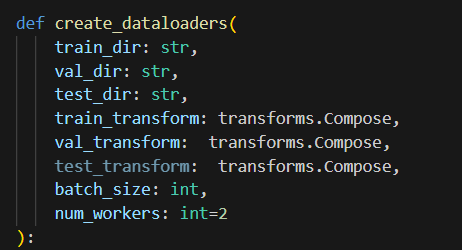
*Image after augmentation*:

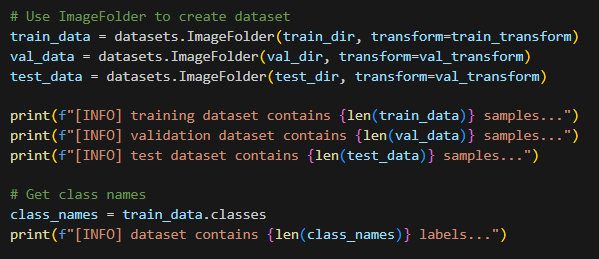


**Implementation**

*Data Loading and Augmentation*

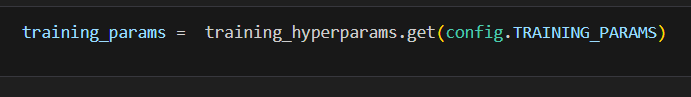
Data loading is managed by PyTorch's DataLoader, which handles batching and shuffling. Augmentations are applied in real-time during loading.

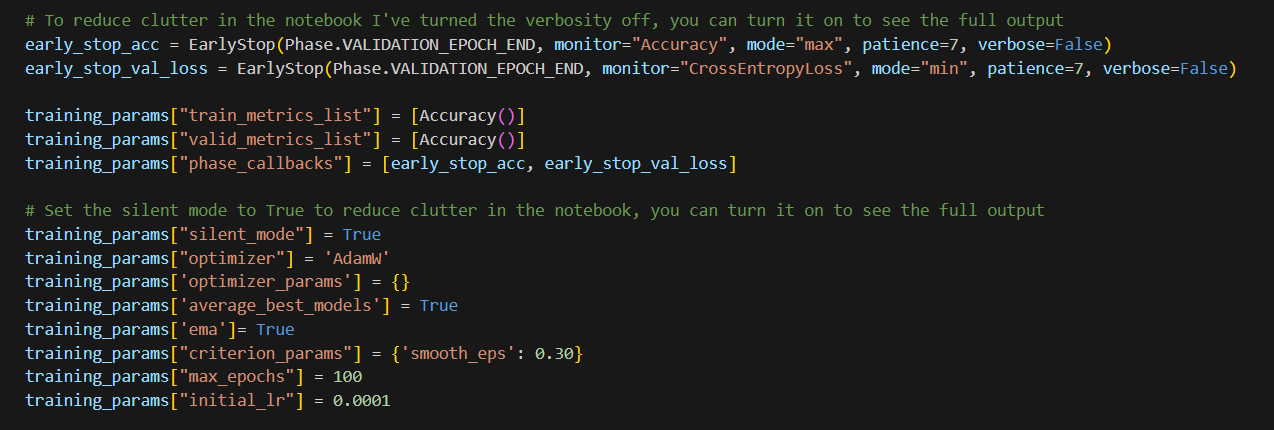




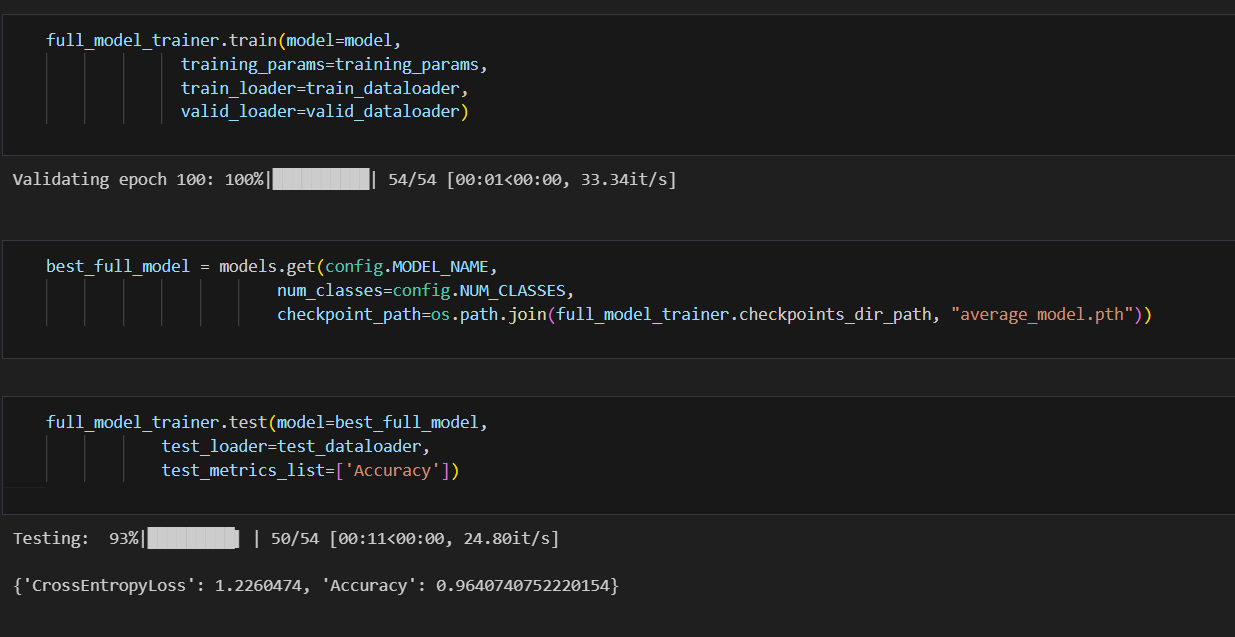
**Model Training**

The ViT model is initialized with pre-trained weights and fine-tuned on the dataset. The Adam optimizer is used to minimize the loss function.



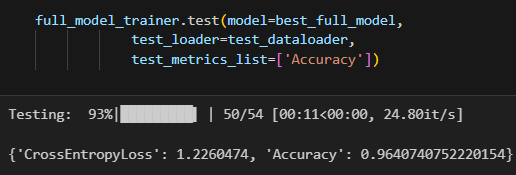
****

Training the model and also get the best model:



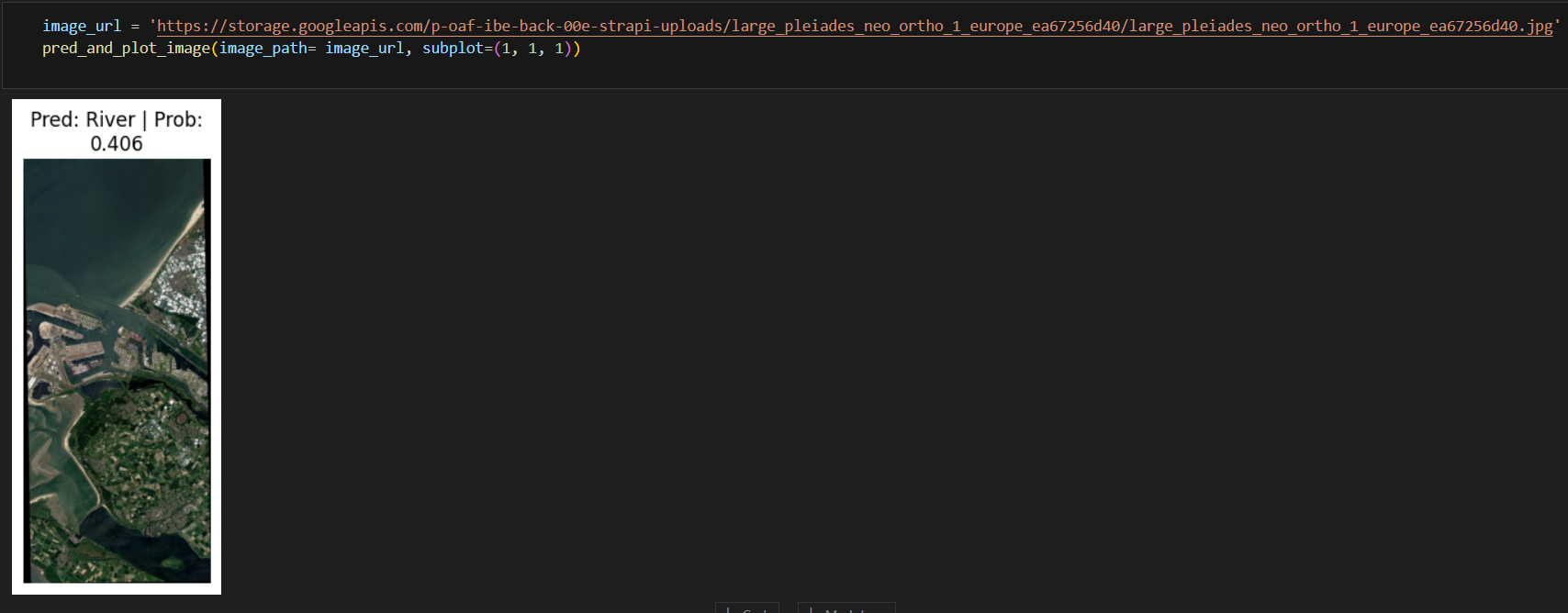
**Evaluation**

Model performance is evaluated using accuracy and other metrics on the validation set.

****

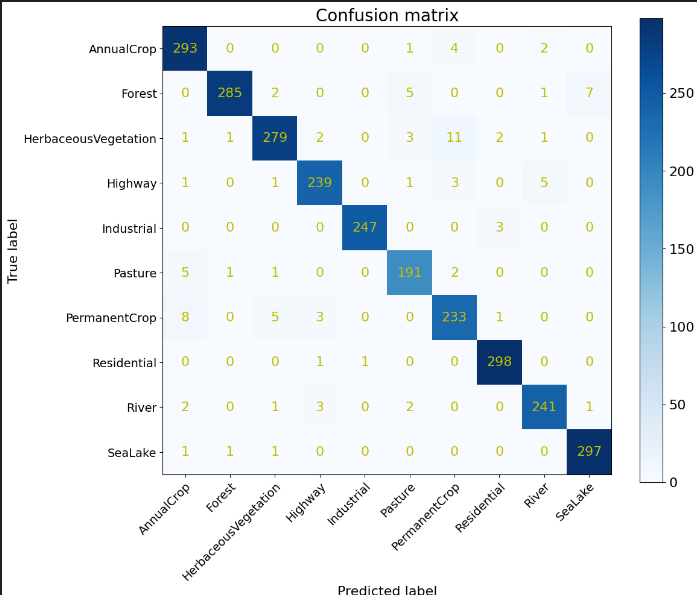
*Plot prediction*

Draw the prediction of some random image and also directly an image from internet:

****

*Confusion Matrix*

This helps assess the accuracy and reliability of the model, and also identify where it makes the most errors.



**Results**

The ViT model achieved high accuracy on the test set, indicating its effectiveness in classifying satellite images. The model's predictions were visualized, showing correct classifications for most images.

**Conclusion**

Vision Transformers are effective for satellite image classification, providing an alternative to traditional CNNs. Future work could explore larger datasets and more extensive hyperparameter tuning to further enhance performance.

**References**

- [Kaggle Notebook: Satellite Image Classification with ViT](https://www.kaggle.com/code/harpdeci/satellite-image-classification-with-vit/notebook?fbclid=IwZXh0bgNhZW0CMTAAAR3AM7cY_VfIwwIxIfEyJyH2ck7zLlahOdw7cTOm0-GhLT9ibAaGzRrDpTM_aem_e6Y_4JxKXU8WaIVqK-aQPg)

- [EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification](https://zenodo.org/records/7711810?fbclid=IwZXh0bgNhZW0CMTAAAR3AM7cY_VfIwwIxIfEyJyH2ck7zLlahOdw7cTOm0-GhLT9ibAaGzRrDpTM_aem_e6Y_4JxKXU8WaIVqK-aQPg#.ZAm3k-zMKEA)