ILEGAL MINING

Authors[1]

[1]Universities

ABSTRACT

This project aims to address global illegal mining detection by developing a solution through fine-tuning the Segment Anything Model (SAM) or integrating lightweight adapters. By adapting SAM's segmentation capabilities to the specific visual signatures of illegal mining activities in diverse geographical contexts, we seek to create a highly accurate and efficient system for monitoring and detection using satellite imagery.

1 Label Generation Pipeline

1.1 Base Mining Mask Generation

The base mining mask is generated from polygon annotations stored in a KML file. For each image, we first identify relevant mining polygons by matching the site name. The polygons are rasterized onto the image grid using the following equation:

$$M_{base}(x, y) = \begin{cases} 2 & \text{if } (x, y) \in \bigcup_i P_i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where P_i represents the set of mining polygons for the site, and (x, y) are pixel coordinates in the image space.

1.2 Vegetation Filtering using NDVI

To remove vegetation areas from the mining mask, we calculate the Normalized Difference Vegetation Index (NDVI) using the near-infrared (NIR, Band 4) and red bands (Band 3). The NDVI is calculated as:

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}} \tag{2}$$

∗ Corresponding author

The mining mask is then updated by setting pixels with NDVI > 0.5 to background (0):

$$M_{veg}(x,y) = \begin{cases} 0 & \text{if NDVI}(x,y) > 0.5 \\ M_{base}(x,y) & \text{otherwise} \end{cases} \tag{3}$$

1.3 Cloud Mask Integration

The final label incorporates cloud information from the 8th band of the original image. From the image statistics, we can see that Band 8 (cloud mask) contains binary values, indicating this is a binary cloud mask. The three-class label is generated with the following priority: cloud (1) > mining (2) > background (0):

$$L(x,y) = \begin{cases} 1 & \text{if } C(x,y) = 1 \\ 2 & \text{if } M_{veg}(x,y) = 2 \text{ and } C(x,y) = 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $C(x,y)$ represents the cloud mask from the 8th band.

1.4 Image Processing

The original Landsat image contains 8 bands, where the 8th band is the cloud mask. For the input images, we only keep the first 7 bands, discarding the cloud mask band since it's already incorporated into the labels. The labels are saved in PNG format with a viridis colormap to ensure clear visualization of the three classes (background, cloud, and mining areas). This pipeline ensures that:

- Mining areas are accurately represented by the polygon annotations
- Vegetation is properly excluded from mining areas
- Cloud coverage is properly labeled with highest priority
- The final labels are easily interpretable with distinct colors for each class

The resulting dataset can be used for training semantic segmentation models to detect mining areas while accounting for cloud coverage and vegetation interference.

2 Dataset Structure and Organization

2.1 Dataset Overview

The Global Mining Dataset is organized into three main splits: training (75.66%), validation (16.55%), and test (7.79%). Each split contains three types of data:

- **Multi-band images** (tif): Original Landsat images with 7 spectral bands
- **Semantic segmentation labels** (png): Three-class labels (background, cloud, mining)
- **RGB visualizations** (png): Color composites for visual inspection

2.2 Data Split Strategy

The dataset is strategically split to evaluate model performance across different scenarios:

Table 1: Site Information Grouped by Data Split and Continent

Split	Continent	Country	Base Site	Images
Train (75.66%)	Africa	Cameroon	Cameroon Kadei River Batouri	201
		Mali	Mali Faleme Upper	271
		Mozambique	Mozambique Manica	365
	Asia	Indonesia	Indonesia Madreng	326
		Indonesia	Indonesia Batang Hari Bedaro	14
		Indonesia	Indonesia West Kalimantan Selimbau	6
		Myanmar	Myanmar Chindwin River Hkamti	857
		Myanmar	Myanmar Chindwin River Ningbyen	775
		Myanmar	Myanmar Namsi Awng	392
		Myanmar	Myanmar Theinkun	386
		Myanmar	Myanmar Kawbyin	319
		Russia	Russia Novotroitsk	80
		Russia	Russia Tumnin	65
		Russia	Russia Koryak Plateau	26
		Russia	Russia Tumnin Tributary	13
		Russia	Russia Edakuy	9
	South America	Peru	Peru La Pampa South	161
		Peru	Peru La Pampa North	67
		Peru	Peru Rio Inambari Channel	44
		Peru	Peru Tournavista	11
Val (16.55%)	Africa	DR Congo	DRC Lindi River Upper	150
		Nigeria	Nigeria Ijesa	163
	Asia	Indonesia	Indonesia Kulu	249
		Myanmar	Myanmar Maw Luu	218
		Russia	Russia Mongolia Border	105
	South America	Peru	Peru Rio Quimiri Downstream	75
Test (7.79%)	Africa	Senegal	Senegal River Gambie	49
		Sierra Leone	Sierra Leone Pampan River Gold Diamond	144
	Asia	Mongolia	Mongolia Gatsuurt	57
		Philippines	Philippines Quiniput Downstream	9
	South America	French Guiana	French Guiana Deux Branches	11
		Venezuela	Venezuela Yapacana South	130
	North America	Nicaragua	Nicaragua Somotillo	52

2.2.1 Train-Validation Split

The training and validation sets are designed with overlapping mining sites to:

- Ensure sufficient data for training complex deep learning models
- Allow validation on similar but distinct mining patterns
- Maintain geographical diversity while keeping some consistency in mining characteristics

2.2.2 Test Set Design

The test set is carefully curated to evaluate model generalization:

- Contains completely different mining sites from training/validation
- Spans across all continents (Africa, Asia, South America, North America)
- Represents real-world scenarios where models need to generalize to unseen locations

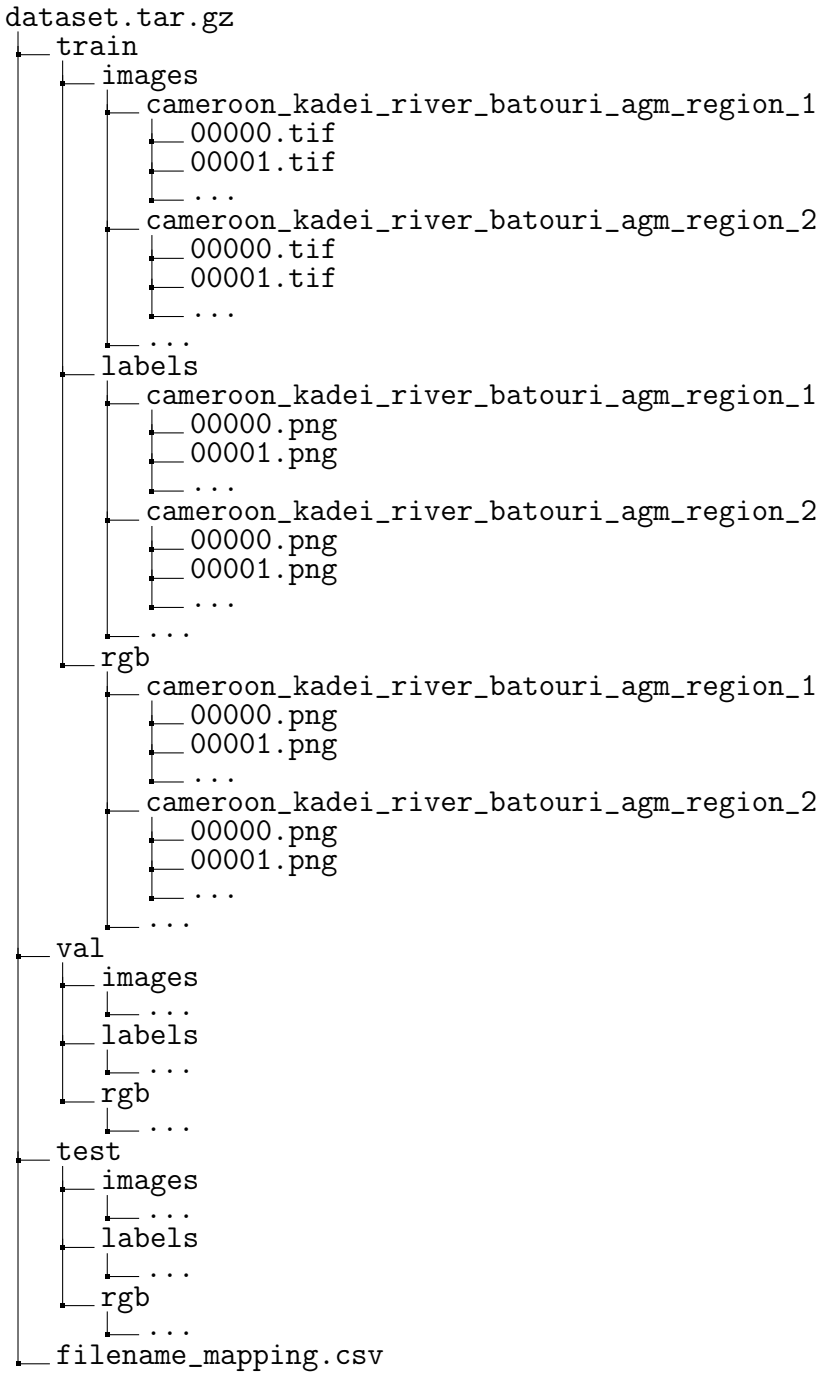

Figure 1: Global Mining Dataset Structure. The dataset contains three splits (train/val/test), each with three types of data: original multi-band images (tif), semantic segmentation labels (png), and RGB visualizations (png). The filename_mapping.csv provides the mapping between original and new filenames. **Download the data here**.

2.3 Geographical Distribution

The dataset covers mining sites across multiple continents and countries:

- **Africa**: Cameroon, Mali, Mozambique, DR Congo, Nigeria, Senegal, Sierra Leone
- **Asia**: Indonesia, Myanmar, Russia, Mongolia, Philippines
- **South America**: Peru, French Guiana, Venezuela
- **North America**: Nicaragua

2.4 Data Organization

Each split follows the same directory structure:

- `images/`: Contains the original multi-band Landsat images
- `labels/`: Contains the three-class semantic segmentation labels
- `rgb/`: Contains RGB visualizations for easy inspection

Files are organized by mining site, with each site having its own subdirectory. The naming convention uses a 5-digit format (e.g., 00000.tif) to maintain consistent ordering. A `filename_mapping.csv` file is provided to track the relationship between original and new filenames.