Project Update: Curating Robust Datasets for Fire and Smoke Detection

Date: November 30, 2025

Subject: Completion of Dataset Curation and Preparation for High-Performance Training

We have successfully completed the most critical and time-intensive phase of this project: the comprehensive acquisition, filtration, and curation of the training and validation datasets. This foundational work ensures maximum model performance and robustness in real-world environments.

1. Data Acquisition and Strategic Curation (High Effort Phase)

The process began with the aggregation of high-quality data from multiple established sources, including:

- aiformankind/wildfire-smoke-dataset
- acquire.cqu.edu.au/articles/dataset/Annotated_Fire_-Smoke_Image_Dataset...
- kaggle.com/datasets/sayedgamal99/smoke-fire-detection-yolo
- etsin.fairdata.fi/dataset/1dce1023-493a-4d63-a906-f2a44f831898/data

The core effort involved transforming these raw sources into **one unified, highly curated dataset** tailored to our primary goal (wildfire detection). This collection was built for maximum robustness:

- **Positive Targets:** It is rich in high-confidence images containing **smoke only** and **fire and smoke**, aligning directly with the core use case.
- **Challenging Negatives:** It includes a significant volume of challenging negative examples, such as heavy thick clouds, red flashes, sunlight glares, yellow non-smoke objects, and plain aerial (UAV) forest images. This critical inclusion prevents false positive detections in complex, real-world scenes.

Dataset Split for Practicality: Due to the large size and constraints related to cloud upload/training management, this unified collection was **split into two equally comprehensive datasets**. Both resulting datasets contain the necessary balance of positive and negative samples for robust training.

2. Enhancing Model Robustness

To ensure the final model is reliable in challenging conditions, significant time was invested in filtration, standardization, and augmentation:

- **Annotation Standardization (High Effort):** We manually converted all bounding box annotations from various formats (e.g., **PASCAL VOC XML**) to the precise, normalized format required by the **YOLO** training pipeline. This included meticulously checking and correcting the class labels to ensure a consistent standard (`0` for Smoke, `1` for Fire) across all sources, as several raw datasets used inverted labels.
- **Comprehensive EDA and Filtration:** We performed an exhaustive Exploratory Data Analysis (EDA) on the aggregated data. This included visualizing the bounding boxes (BBox) on the images to meticulously verify that the annotation conversions and dataset splitting were perfectly aligned, ensuring no inaccuracies would compromise training. This process also led to the removal of false positives like cigarette smoke, steam, and other irrelevant forms of vapor.
- **Night-Time Robustness:** Recognizing a critical gap in publicly available data, we introduced several strategic augmentation techniques—specifically adjusting **HSV-V (Value/Brightness), Mosaic, and Close Mosaic**—to simulate diverse night-time conditions. This is essential for $24/7$ operational reliability.

3. Current Status

The finalized, high-quality datasets (in two parts) have been compiled, zipped, and uploaded to the cloud service, ready for the training phase. The quality of this data significantly reduces the risk of model failure and ensures high accuracy.

Current Bottleneck:

The sole remaining dependency is the configuration and allocation of the proper high-performance GPU environment on the cloud. The sheer volume and complexity of the dataset (even after image size optimization) require substantial computational power to begin training and experimentation efficiently. Once this environment is active, the model training phase will commence immediately.

The data curation is the single most time-intensive and valuable step in this process. The quality of the foundation is now secure.