dataset_instruction

Dataset Setup Instructions

> **The dataset is NOT included in this repository** (it's ~2GB+ and that's bad practice). Follow these steps to download and place the files correctly.

Step 1 — Create a Falcon Account (Free)

1. Go to Falcon Sign Up
2. Create a free account with your email
3. Verify your email and log in

Step 2 — Download the Dataset

1. Go to the Hackathon Dataset Page
2. You will see **3 download links** — download ALL of them:

#	Download	What it contains	Approx. Size
1	**Offroad_Segmentation_Training_Dataset**	Training + validation images and masks	~1.5 GB
2	**Offroad_Segmentation_testImages**	Unseen test images (RGB only, no masks)	~200 MB
3	**Offroad_Segmentation_Scripts**	Starter training/testing scripts + env setup	~10 MB

3. Each download will be a `.zip` file — extract them all.

Step 3 — Place Files in the Correct Location

After extracting, place the 3 folders inside the `DATASET/` directory so the structure looks like this:

```
DATASET/
├── Offroad_Segmentation_Training_Dataset/
│   ├── train/
│   │   ├── Color_Images/          ← 2857 RGB images (960×540, PNG)
│   │   └── Segmentation/          ← 2857 mask images (uint16, PNG)
│   └── val/
│       ├── Color_Images/          ← 317 RGB images
│       └── Segmentation/          ← 317 mask images
│
├── Offroad_Segmentation_testImages/
│   └── *.png                      ← ~500 test RGB images (no masks!)
│
├── Offroad_Segmentation_Scripts/  ← Starter scripts from hackathon
│   ├── train_segmentation.py
│   ├── test_segmentation.py
│   ├── visualize.py
│   └── ENV_SETUP/
│
└── dataset_instruction.md         ← You are here
```

Step 4 — Verify the Setup

Run this quick check to make sure everything is in place:

Windows (PowerShell):

```
# Count training images
(Get-ChildItem
"DATASET\Offroad_Segmentation_Training_Dataset\train\Color_Images" -Filter
"*.png").Count
# Expected: 2857

# Count validation images
(Get-ChildItem
"DATASET\Offroad_Segmentation_Training_Dataset\val\Color_Images" -Filter
"*.png").Count
# Expected: 317

# Count test images
(Get-ChildItem "DATASET\Offroad_Segmentation_testImages" -Filter
"*.png").Count
# Expected: ~500
```

Linux/macOS:

```
# Count training images
ls DATASET/Offroad_Segmentation_Training_Dataset/train/Color_Images/*.png | wc -l
# Expected: 2857

# Count validation images
ls DATASET/Offroad_Segmentation_Training_Dataset/val/Color_Images/*.png | wc -l
# Expected: 317

# Count test images
ls DATASET/Offroad_Segmentation_testImages/*.png | wc -l
# Expected: ~500
```

Image Format Details

Color Images (RGB)

- **Format**: PNG, 3-channel RGB
- **Resolution**: 960 × 540 pixels
- **Content**: Synthetic desert scenes rendered by Falcon/Duality AI

Segmentation Masks

- **Format**: PNG, single-channel uint16
- **Resolution**: 960 × 540 pixels (same as RGB)
- **Pixel values**: Raw class IDs that need mapping:

Pixel Value	Class ID	Class Name
0	0	Background
100	1	Trees
200	2	Lush Bushes
300	3	Dry Grass
500	4	Dry Bushes

Pixel Value	Class ID	Class Name
550	5	Ground Clutter
700	6	Logs
800	7	Rocks
7100	8	Landscape
10000	9	Sky

> ⚠️ **Important**: Masks must be read as uint16 (not uint8!) because values go up to 10000. Use `cv2.imread(path, cv2.IMREAD_UNCHANGED)` or `PIL.Image.open(path)` to preserve the full value range.

Class Distribution (Training Set)

Class	Pixel %	Rarity
Sky	34.72%	Very common
Landscape	22.34%	Common
Dry Grass	17.37%	Common
Lush Bushes	5.50%	Moderate
Background	5.36%	Moderate
Ground Clutter	4.03%	Moderate
Trees	3.29%	Uncommon
Rocks	1.10%	Rare
Dry Bushes	1.01%	Rare
Logs	**0.07%**	**Extremely rare**

> Sky is **500× more common** than Logs. Class weighting or Focal Loss is essential for fair training.

Rules (from Hackathon Guidelines)

1. ✅ Train **only** on the provided dataset — no external data allowed

2. ❌ **NEVER** use test images for training or validation — instant disqualification
3. ✅ You may use any augmentations, architectures, or training strategies
4. ✅ Pre-trained backbones (like DINOv2) are allowed