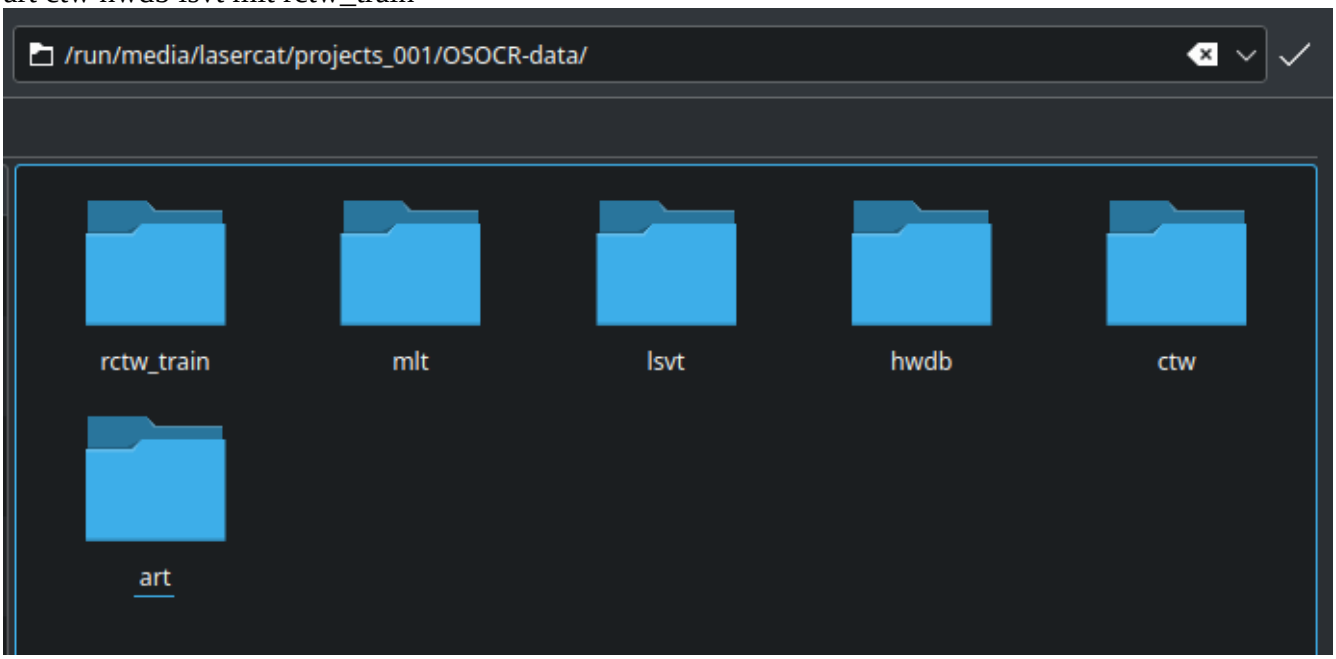
Data Builder for OSOCR-Family

1. Building datasets used in the paper.

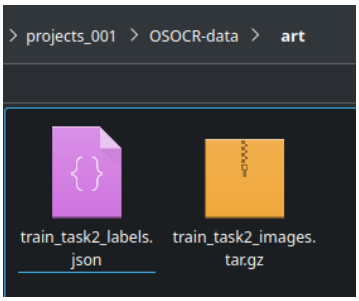
1. Make a dataset source dir ${SRC}, which has the following subfolders:

art ctw hwdb lsvt mlt rctw_train



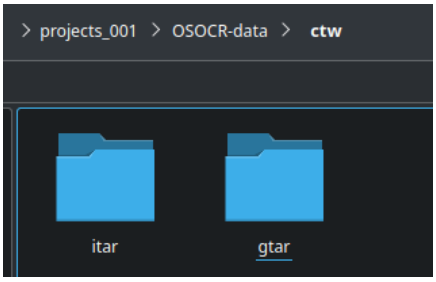
2. Download datasets listed in dataset_sources.txt into corresponding dirs. Since I am on a paid-by-data network, let me skip the re-downloading process… If anything goes south, please open an issue.

2.1 Art:
After downloading, make sure the following two files lies in the art folder:
train_task2_images.tar.gz and train_task2_labels.json

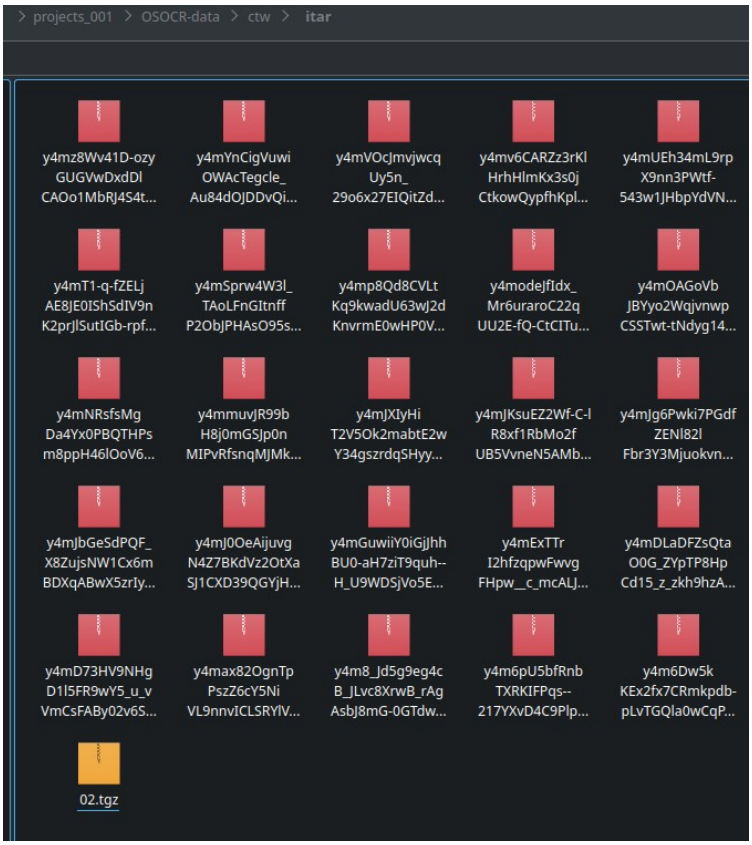


2.2 CTW:

2.2.1 Make two folders: itar for image, and gtar for gts.

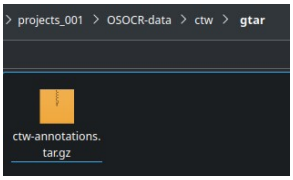


2.2.2 Download all parts into itar

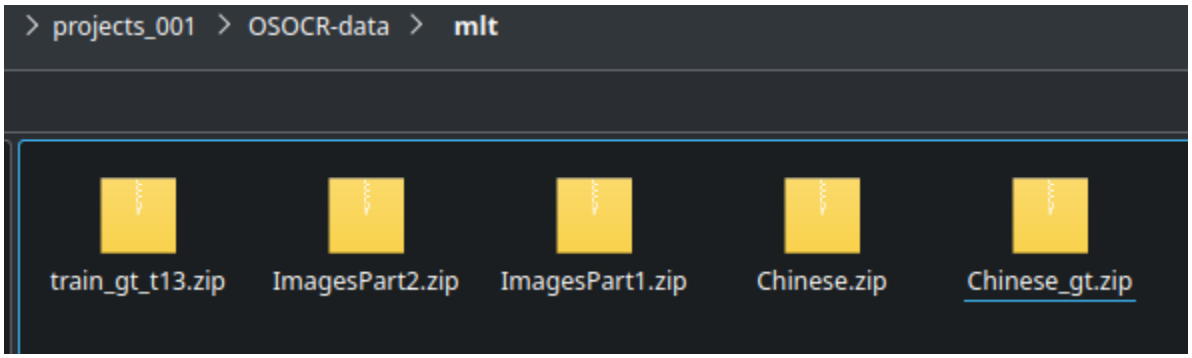


2.2.3. Make a tarlist for the CTW dataset:
ls |grep -v tarlist>tarlist

2.2.4. Download annotation to the gtar folder:

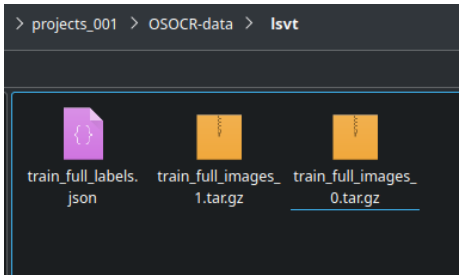


2.3. Download MLT-19 real and Chinese synthetic data to the mlt folder.

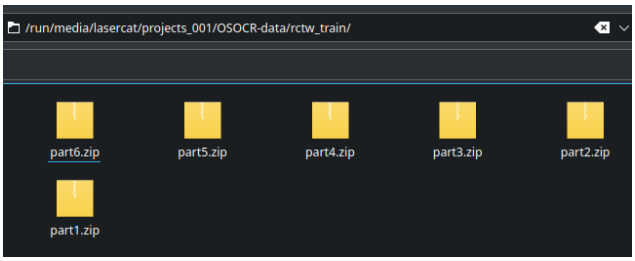


The synthetic data is not actually used…. But… Please make sure it's there...

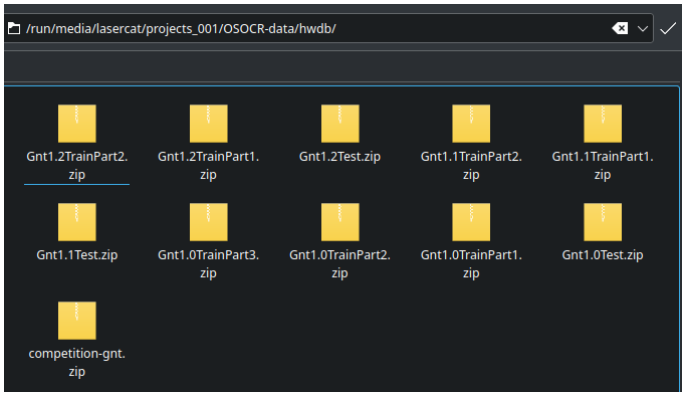
2.4. Download the LSVT data to the lsvt folder.
Only the fully supervised part needed.



2.5 Download the 6 parts of rctw training set into the rctw_train folder.



2.6 Download OLHWDB 1.0-1.2 and the competition set to the hwdb folder:



2.7 Almost there! Let’s check the recipe:

ls -R >all.my;
Compare all.my to the all.txt in the repo, make sure you do not miss a file or two

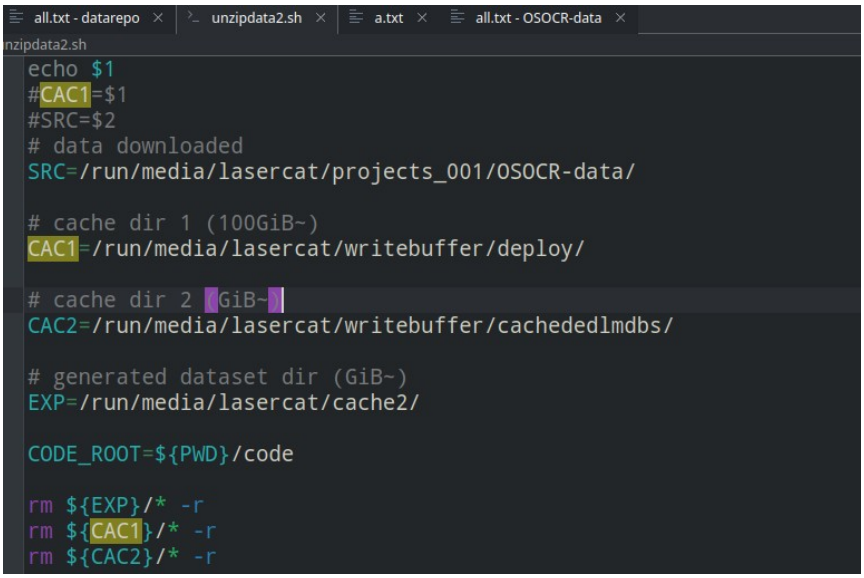
3. Set up enviornment

The scripts contain some rmdir, rm, mv commands, so please isolate it from important data.
!!!!YOU ARE WARNED!!!!!

3.1 Follow this link to setup
https://github.com/lancercat/make_env/

3.2 Buy a used 240Gib SSD from ebay or elsevier. Two 120 Gib drives shall do as well. The gist is not to write a lot temporary files to your expensive main SSD.

3.3 Setup paths:
Open up unzipdata2.sh and set SRC, CAC1, CAC2, and EXP.



SRC is your dataset source folder.
CAC1 is one cache folder, and CAC2 is another, redirect them to some cheap disks. EXP is where built datasets are going to be stored. These folders should be **EMPTY**
The scripts contain some rmdir, rm, mv commands, so please CORRECTLY set the paths.
One step wrong, your data GONE.
!!!!YOU ARE WARNED!!!!!
4. Grab some snacks and push the **RED BUTTON**, wait for a few hours and your lmdbs will be there:
sh unzipdata2.sh

2. Building custom datasets.