Noisy Labels repairing in dataset CASIA v2.0 groundtruth

Intro

CASIA 2.0 is a dataset for Image Tampering Detection Evaluation, which was published by Jing Dong et al in 2013. However, this dataset is **lack of the groundtruth images** comparing to other Image Tampering Detection Datasets.

To soleving the problem, Nam Thanh Pham et al. generated the corresponding Groundtruth in a 2019 paper contributed it to Github. This publicly available groundtruth has gained wide distribution in data science platforms such as kaggle

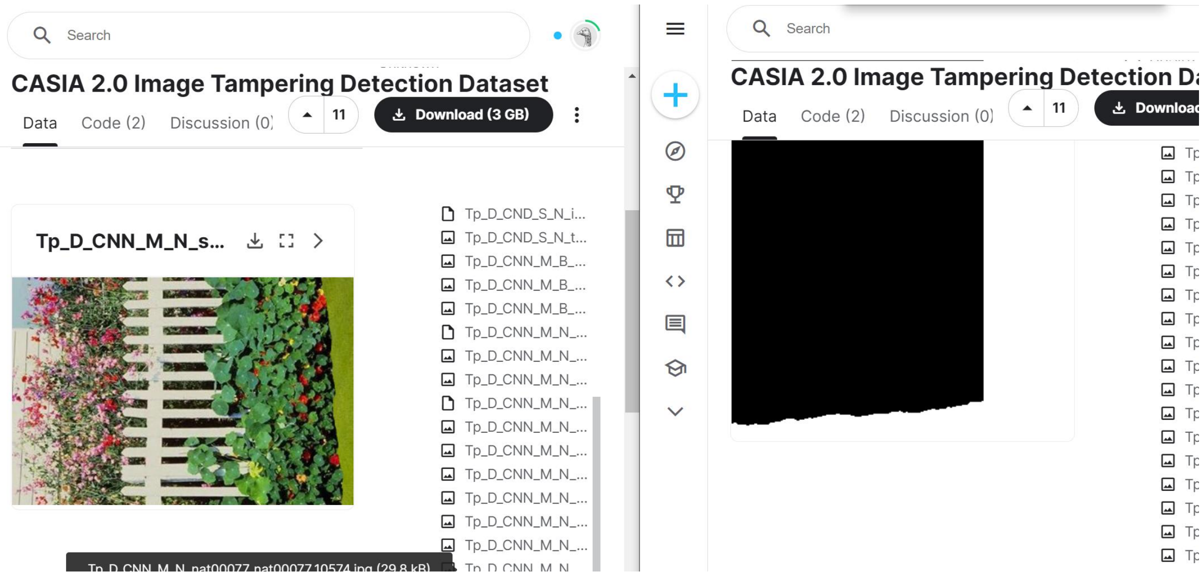
> Nam Thanh Pham et al. also corrected some mistakes in naming the files of the origianl CAISA 2.0 Dataset.

Noisy Labels in wide spread groundtruth

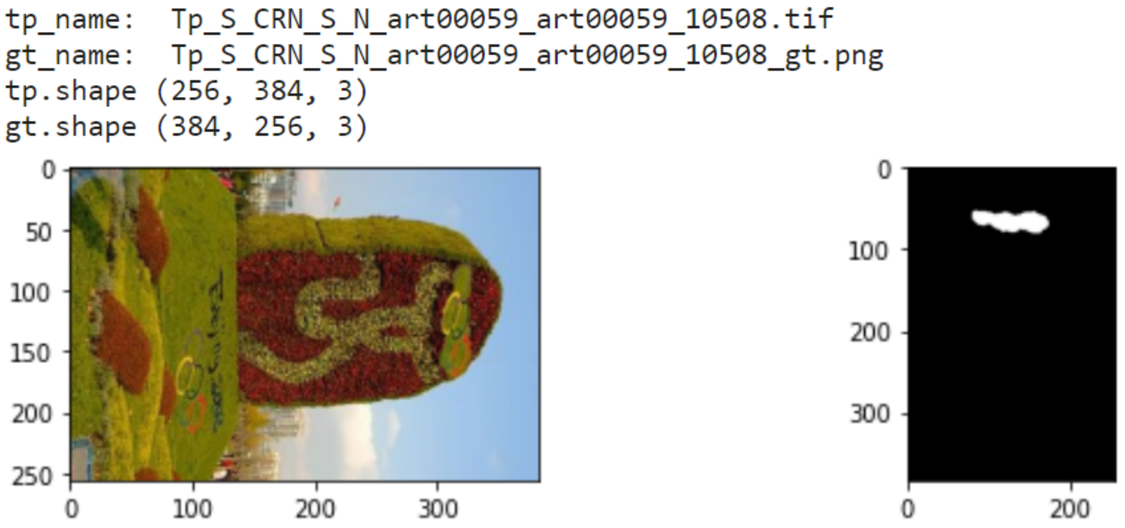
However, when we doing experiments base on CASIA 2.0 datasets, we found that there are some **serious noises** in groundtruth such as :

- Rotation mismatch
- Resolution mismatch
- Mask boundary mismatch

Here are some Examples:

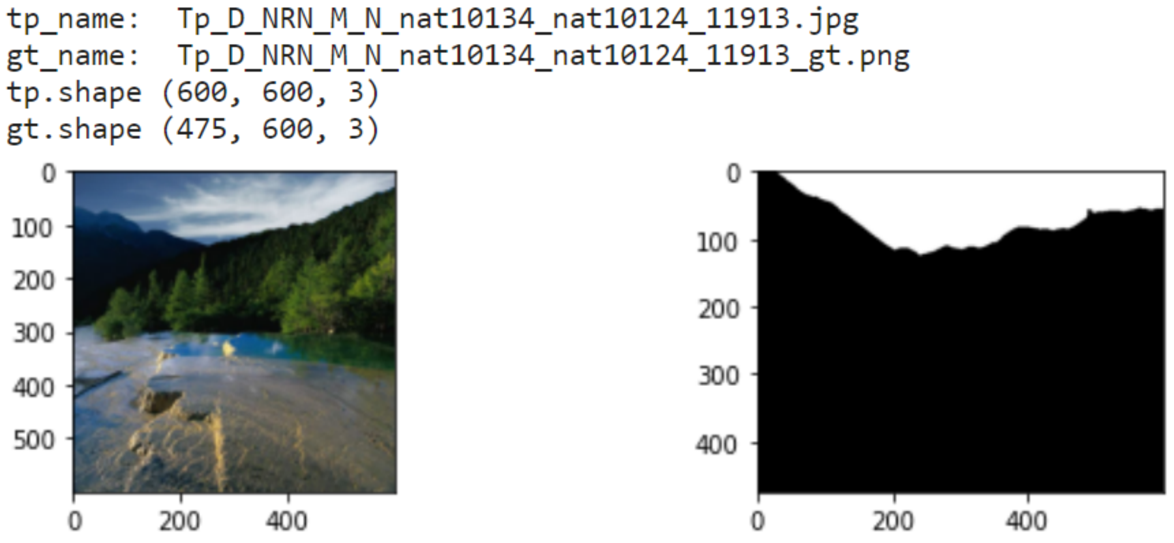
- Rotation mismatch example on Kaggle platform:
  - 

- Another Rotation mismatch example:



```
tp_name:  Tp_S_CRN_S_N_art00059_art00059_10508.tif
gt_name:  Tp_S_CRN_S_N_art00059_art00059_10508_gt.png
tp.shape (256, 384, 3)
gt.shape (384, 256, 3)
```

- Resolution mismatch example:



```
tp_name:  Tp_D_NRN_M_N_nat10134_nat10124_11913.jpg
gt_name:  Tp_D_NRN_M_N_nat10134_nat10124_11913_gt.png
tp.shape (600, 600, 3)
gt.shape (475, 600, 3)
```

Because `resize()` is generally used in pre-processing, these dozens of problematic images are difficult to be detected from more than 5000 tampered images.

What's more, this dataset is widly use in the field of Image Tampering Detection to eavaluate model performance, and it's hard to find a second groundtruth dataset on the Internet, we have reason to believe that many papers have adopted this groundtruth as the validation of the CASIA 2.0 dataset.

Fixed groundtruth downloading

Although these images can hardly have a significant impact on the training results of a dataset containing more than 5,000 images, we thought it would be useful to point out this issue for researchers to know.

And here we place the Google Drive link of corrected CASIA 2.0 dataset and it's ground truth

Cite