

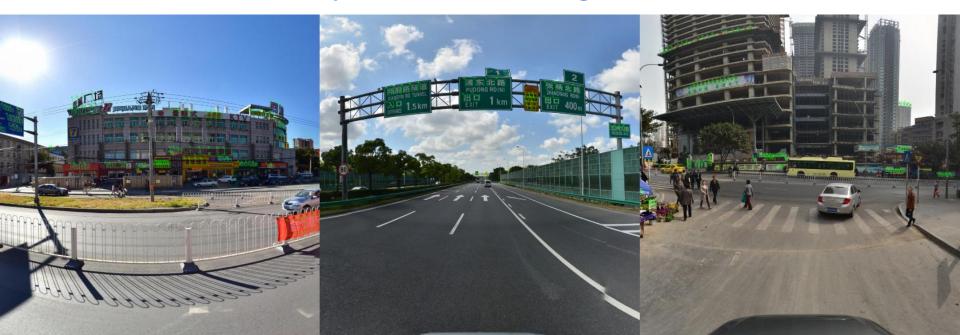
A Large Chinese Text Dataset in the Wild

Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu

Yuan TL, Zhu Z, Xu K et al. A large Chinese text dataset in the wild. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 34(3): 509–521 May 2019. DOI 10.1007/s11390-019-1923-y

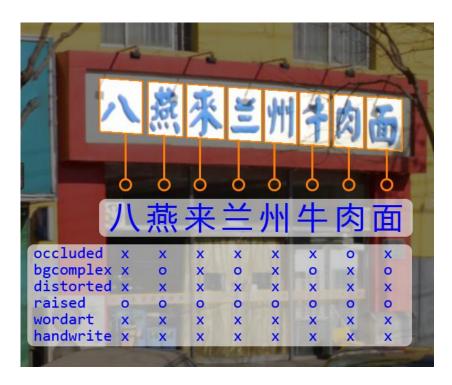
Dataset Glance

- 32,285 high resolution images
- **1,018,402** character instances
- 3,850 character categories
 - https://ctwdataset.github.io/



Annotation Format

- underlying character
- bounding box
- 6 kinds of attributes
 - occluded / not occluded
 - complex / clean bg.
 - distorted / frontal
 - 3D raised / planar
 - wordart / not wordart
 - handwritten / printed



Text Detection Task

Character detection baseline

Text line detection baseline



(wrong detections are shown in yellow)