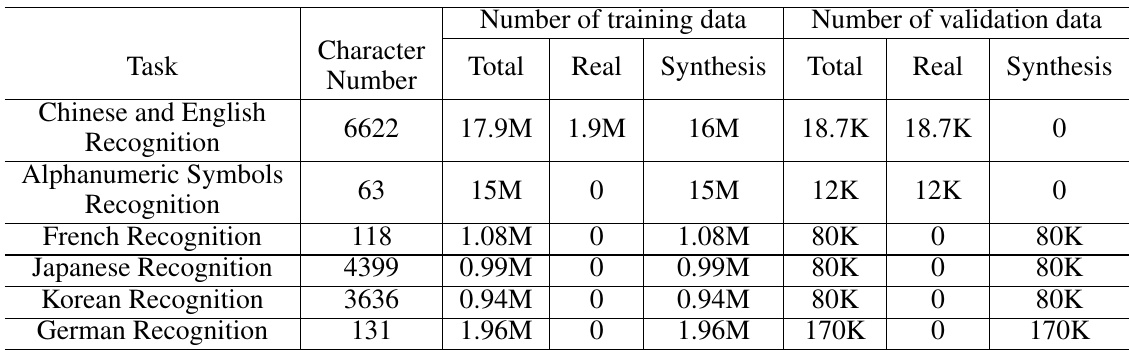


Table 1: Statistics of dataset for Chinese and English Recognition.



For direction classification, there are 600k training images and 310K validation images. Among the training images, 100K images are real scene images, which come from the public datasets (LSVT, RCTW-17, MTWI 2018). They are horizontal text which rectify and crop the ground truth of the images. The remaining 500K synthetic images mainly focus on the reversed text. We use the vertical fonts to synthesize some text images and then rotate them horizontally. All the validation images come from the real scenes.

validation images come from the real scenes. For text recognition, there are 17.9M training images and 18.7K validation images. Among the training images, 1.9M images are real scene images, which come from some pub- lic datasets and Baidu image search. The public datasets used include LSVT, RCTW-17, MTWI 2018 and CCPD 2019. The remaining 16M synthetic images mainly focus on the scenarios for different backgrounds, translation, rotation, perspective transformation, line disturb, noise, vertical text and so on. The corpus of synthetic images come from the real scene images. All the validation images also come from the real scenes. In order to conduct ablation experiments quickly and

In order to conduct ablation experiments quickly and choose the appropriate strategies, we select 4k images from the real scene training images for text detection, and 300k ones from the real scene training images for text recogni- tion.

In addition, we collected 300 images for different real ap. plication scenarios to evaluate the overall OCR system, in- cluding contract samples, license plates, nameplates, train tickets, test sheets, forms, certificates, street view images, business cards, digital meter, etc. Figure[3and Figure4show some images of the test set.

Furthermore, to verify the proposed PP-OCR for other languages, we also collect some corpus for alphanumeric symbols recognition, French recognition, Korean recogni- tion, Japanese recognition and German recognition. Then

Table 2: Statistics of dataset for multilingual recognition.

synthesize the text line images for text recognition. Some images for alphanumeric symbols recognition come from the public datasets, ST (Gupta, Vedaldi, and Zisserman 2016) and SRN (Yu et al. 2020). Table2shows the statistics. Since MLT 2019 for text detection includes multilingual im- ages, the text detector for Chinese and English recognition also can support multi language text detection. Due to the limited data, we haven't found the proper data to train the direction classifier for multilingual.

direction classifier for multilingual. The data synthesis tool used in text detection and text recognition is modified from text render (Sanster[2018).

Implementation Details We use Adam optimizer to train all the models and adopt cosine learning rate decay as the learning rate schedule. The initial learning rate, batch size and the number of epochs for different tasks can be found in Table When we obtain the trained models, FPGM pruner and PACT quantization can be used to reduce the model size further with the above models as the pre-trained ones. The training processes of FPGM pruner and PACT quantization are similar as previous.

In the inference period, HMean is used to evaluate the per- formance of a text detector. Accuracy is used to evaluate the performance of a direction classifier or a text recognizer. F- score is used to evaluate the performance of an OCR system. In order to calculate F-score, a correct text recognition result should be the accurate location and the same text. GPU in- ference time is tested on a single T4 GPU. CPU inference time is tested on a Intel(R) Xeon(R) Gold 6148. We use the Snapdragon 855 (SD 855) to evaluate the inference time of the quantification models.

# 3.2Text Detection

Table 5] compares the performance of the different back- bones for text detection. HMean, the model size and the in-