Supplementary Materials: RDLNet: A Novel and Accurate Real-world Document Localization Method

Anonymous Authors

1 MORE IMPLEMENTATION DETAILS

By default, during distillation process, we take advantage of the original SAM consisting of a ViT image encoder, a Prompt Encoder and a Mask Decoder with layer settings shown in Table 1 for teacher model.

Table 1: SAM Teacher Model Settings

Item	Value
embedding dimension	1280
depth	32
number of heads	16
global attention indexes	[7, 15, 23, 31]
prompt embedding dimension	256
image size	1024
patch size	16

Considering the fact that mobile devices bear limited computational resources, we design the lightweight student model with layer settings shown in Table 2 for student model with a smaller embedding dimension, depth, number of heads yet the same prompt embedding dimension, patch size as the teacher model. During distillation, prompt encoder and mask decoder are frozen and solely the image encoder is trained since we only require a well-trained and generalizable backbone for downstream document localization task.

Table 2: SAM Student Model Settings

Item	Value
embedding dimension	384
depth	12
number of heads	8
global attention indexes	[2,8]
prompt embedding dimension	256
image size	1024
patch size	16

For RDLNet structural hyper-parameters, we adopt the setting shown in Table 3

2 EFFICIENCY OF SAM AND RDLNET ON DOCUMENT LOCALIZATION

We compare the time efficiency of SAM and RDLNet on document localization task. The comparison is shown in Table 4. The inference time is measured on a single NVIDIA A800 for both methods. Averaging time result from evaluating both methods on SmartDoc dataset and Extended SmartDoc dataset, the inference time of SAM is 0.0997 second per image while that of RDLNet is 0.0396 second per image and 2.5 times faster than SAM. On account of the lightweight designs of the model with fewer parameters and GFLOPS, RDLNet achieves decent performance compared to SAM on document localization task, whose efficiency makes it suitable for document localization on mobile devices.

Table 3: RDLNet Structural Hyper-parameters

Item	Value
max image size	1024
number of points embedding	18
number of encoder layers	6
feed forward dimension	2048
number of heads	8
hidden dimension	256
number of feature level	4
number of object queries	5
number of classes	3

Table 4: Comparison of Model Size and Efficiency between RDLNet and SAM

Model	params	GFLOPS	Inference Time
RDLNet	20.55M	100.26G	0.0396
SAM	90.49M	371.98G	0.0997

Table 5: SAM's Performance on SmartDoc dataset and Extended SmartDoc dataset

Dataset	Average JI
SmartDoc	0.5948
Extended SmartDoc	0.6058

3 QUALITATIVE ANALYSIS ON SAM

To reveal the effectiveness of SAM on document localization task, we also conduct exclusive qualitative analysis on SAM using point prompts with experiments on SmartDoc dataset and Extended SmartDoc dataset as shown in Table 5. We make use of point prompts randomly sampled from ground truth annotation mask area to prompt SAM and visualize the prediction output of SAM on SmartDoc dataset and Extended SmartDoc dataset as shown in Figure 1. Nevertheless, both the metric and visualization results manifests that SAM tends to predict fine-grained document areas such as text lines, text blocks, profile pictures and tables closest to the position of user point prompt which is not compatible with the document localization task of detecting the whole document area.

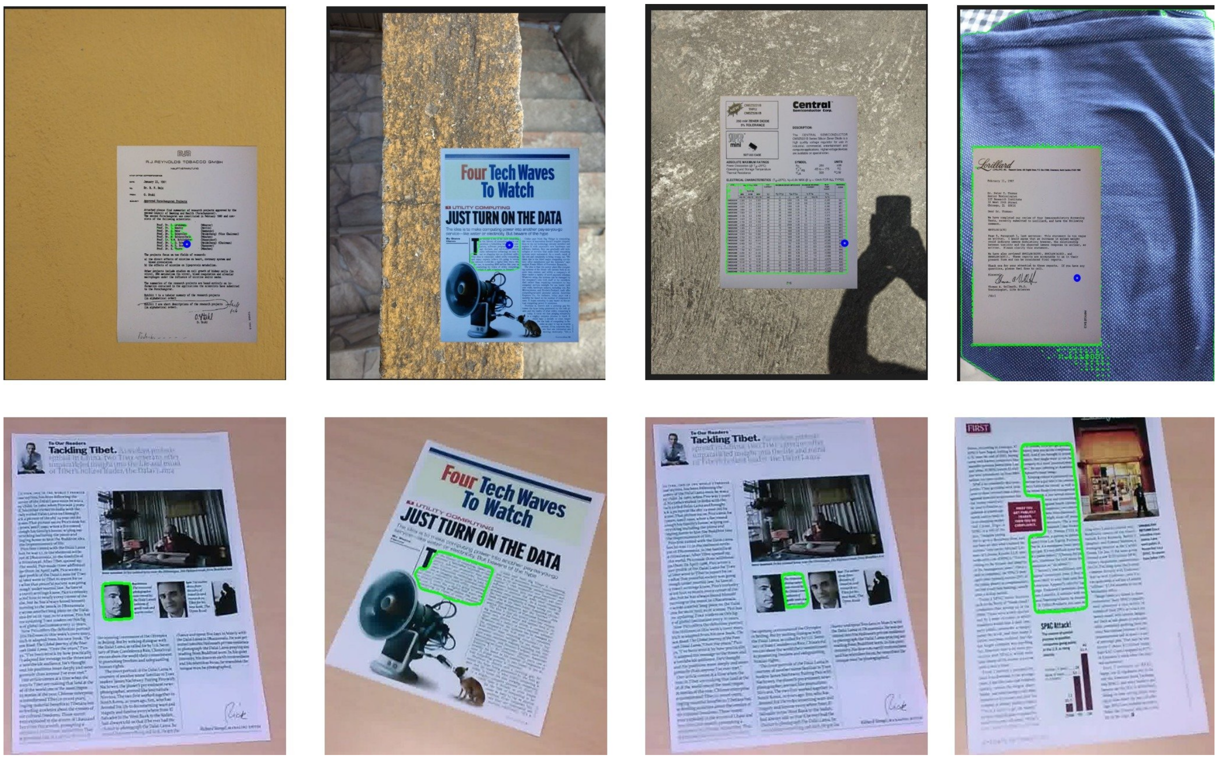


Figure 1: Visualization of SAM prediction output on SmartDoc dataset and Extended SmartDoc dataset

Thus, as a result, SAM itself as a unit is not suitable for document localization, but it possesses affluent semantic knowledge in its image backbone that's worthwhile for transferring to downstream tasks.

4 MORE VISUALIZATIONS OF RDLNET

To better demonstrate the effectiveness of RDLNet, we provide more visualizations of RDLNet's prediction output on SmartDoc dataset and Extended SmartDoc dataset in Figure 2.

The visualizations demonstrate that RDLNet is capable of accurately localizing the whole document area with high precision and recall, which is consistent with the quantitative results in the main paper.

5 SAMPLES OF DIFFERENT CATEGORIES IN RWMD DATASET

In order to more intuitively demonstrate the richness of of RWMD dataset, we present more samples of nine different categories of RWMD dataset in Figure 3. The samples contain various scenarios such as complex background, overlapping documents, occlusion, distortion, shadow, overexposure, low light, low contrast, etc. These samples demonstrate RWMD dataset is a comprehensive and challenging dataset for document localization. We believe that RWMD dataset will prove invaluable for the advancement of the field of document analysis, as well as serve as a benchmark for document localization systems.

6 CODE

Our code will be made publicly available upon camera-ready version.

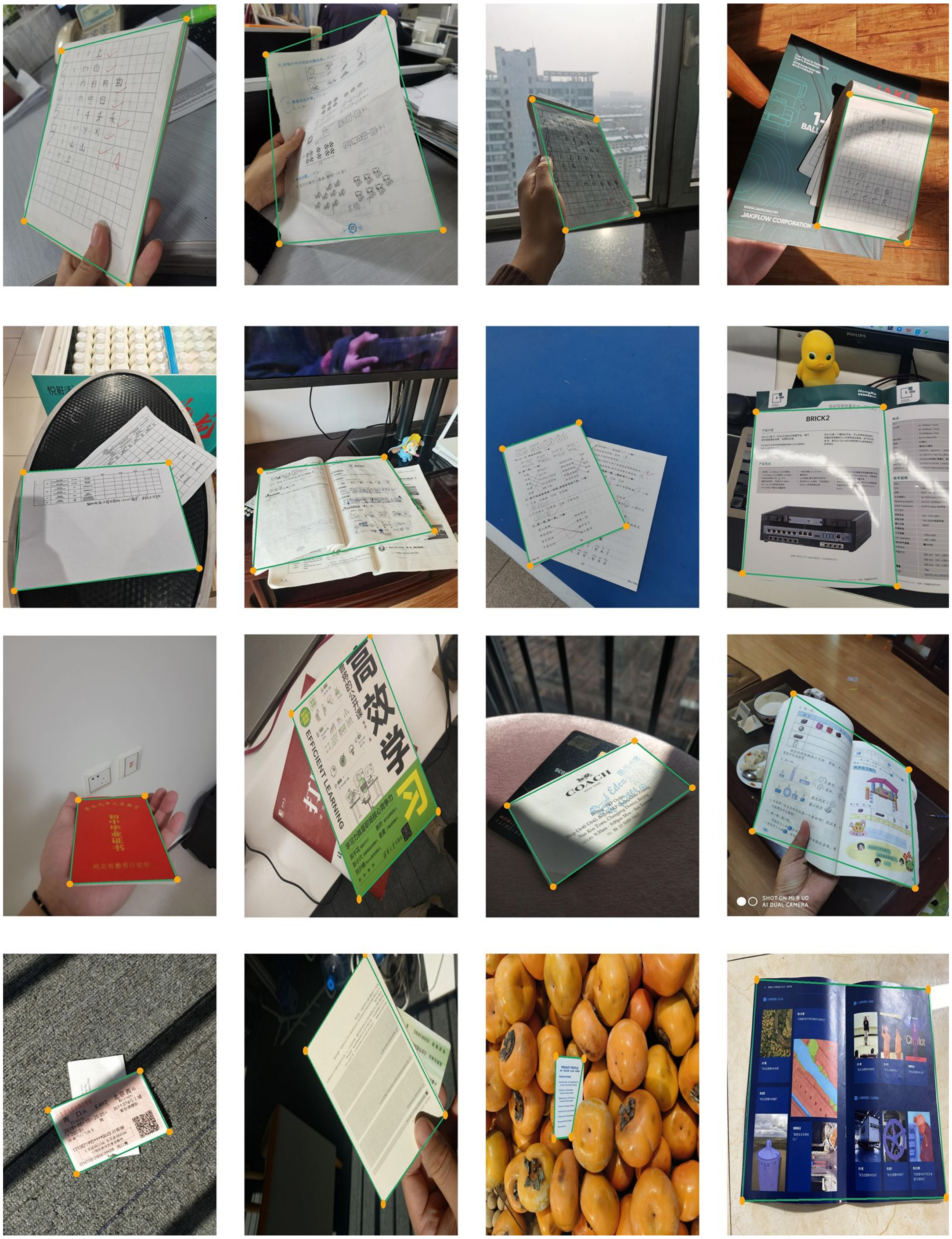
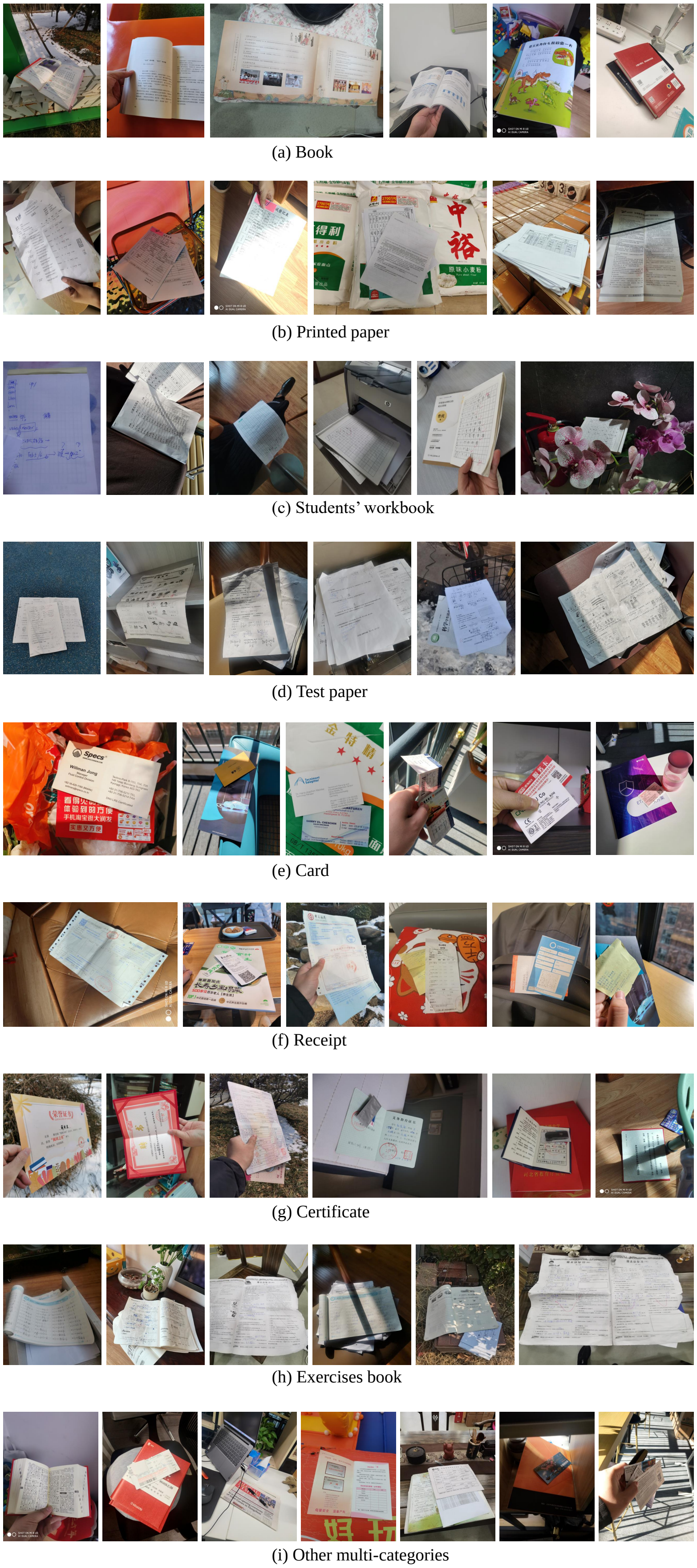


Figure 2: More Visualization of RDLNet prediction output on RWMD dataset



(a) Book

(b) Printed paper

(c) Students’ workbook

(d) Test paper

(e) Card

(f) Receipt

(g) Certificate

(h) Exercises book

(i) Other multi-categories

Figure 3: Samples of different categories in RWMD dataset