From Two to One: A New Scene Text Recognizer with Visual Language Modeling Network (Supplementary Materials)

Yuxin Wang[1], Hongtao Xie[1], Shancheng Fang[1*], Jing Wang[2], Shenggao Zhu[2] and Yongdong Zhang[1]
[1]University of Science and Technology of China
[2]Huawei Cloud & AI
wangyx58@mail.ustc.edu.cn, {htxie,fangsc,zhyd73}@ustc.edu.cn
{wangjing105,zhushenggao}@huawei.com

1. The Evaluation of MLM in Localization

As it is difficult to generate the accurate character-level mask label in the real images, we quantitatively analyze the effectiveness of MLM in character localization by evaluating the character-level recognition in the word image. Specially, we sequentially increase the character index P in MLM from 1 to N (N is the word length predicted by VRM), and combine these character-wise predictions from the first path in MLM (containing only the occluded character) as the final word-level recognition result. For fair comparison, we compare our method with ASTER [2], which uses a similar attention map to focus on the character-wise visual cues in each time step. As shown in Tab. 1, the proposed method obtains better results and outperforms [2] on IC15 dataset by 2.9%, which proves the effectiveness of MLM in locating character-wise visual cues.

We visualize more examples of $Mask_c$ to demonstrate the effectiveness of our MLM in localizing character-wise visual cues. As shown in Fig. 1, the generated $Mask_c$ effectively occludes character-wise visual cues at corresponding position with the guidance of character index P. Specially, MLM can handle the word images with small character spacing (*e.g.* word ”jostle” with $P = 3$, word ”windrow” with $P = 5$, word ”munchkin” with $P = 6$, etc.), the localization of repeated characters (*e.g.* word ”boggles” with $P = 3$, word ”batten” with $P = 4$, word ”pzazz” with $P = 4$, etc.) and the distorted images (*e.g.* word ”mandates” with $P = 3$, word ”dogtrot” with $P = 4$, word ”redcoat” with $P = 4$, etc.). It is worth mentioning that MLM only uses original word-level annotations in the training stage through the proposed *Weakly-supervised Complementary Learning*, accurately localizing character-wise visual cues without the need of additional annotations.

Table 1: The comparisons of the performance in character-level recognition.

Methods	IC13	IC15	SVT	SVTP
ASTER [2]	98.1	90.2	95.7	91.3
MLM	**98.7**	**93.1**	**97.0**	**92.6**

Table 2: The results on ID card dataset.

Methods	Accuracy
Baseline + RNN [2]	95.6
VisionLAN	**97.8**

2. Details of OST Dataset

The Occlusion Scene Text (OST) is a new dataset proposed in this paper, which contains 4832 images selected from 6 benchmarks (IC13, IC15, IIIT5K, SVT, SVTP and CT). The OST dataset is collected considering the diversity of different text shape (*e.g.* horizontal, curved and oriented), text fonts and backgrounds. As the size of this dataset is much larger than most existing scene text recognition datasets, the conclusion reflected on this dataset is more convincing and more general.

As most images in existing benchmarks have clear visual cues of characters, on the case of missing character-wise visual cues, the ability of different methods to use linguistic information for assisting recognition cannot be fully reflected. Since partial texture loss of text is common in scene images (*e.g.* scratch, fading, etc.), we adopt the most straightforward approach to simulate this phenomenon that using lines to occlude characters. Specifically, we manually occlude the images of OST in weak and heavy degrees (shown in Fig. 2 and Fig. 3). Weak and heavy degrees mean that we occlude the character in a text image using one or two lines. For each image, we randomly choose one degree to only cover one character. The color of lines depends on the surrounding background.

[*] Corresponding author

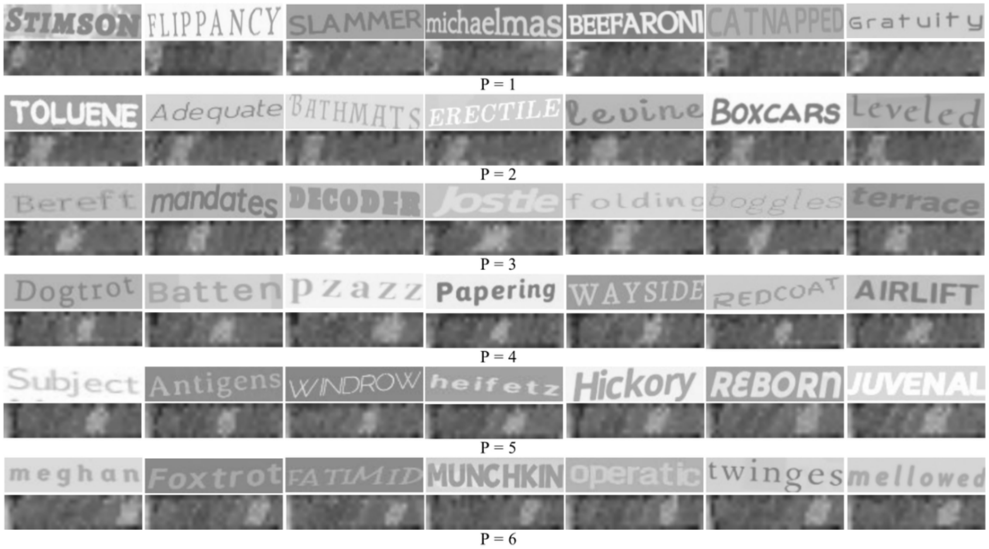

Figure 1: The examples of $Mask_c$. P is the character index.

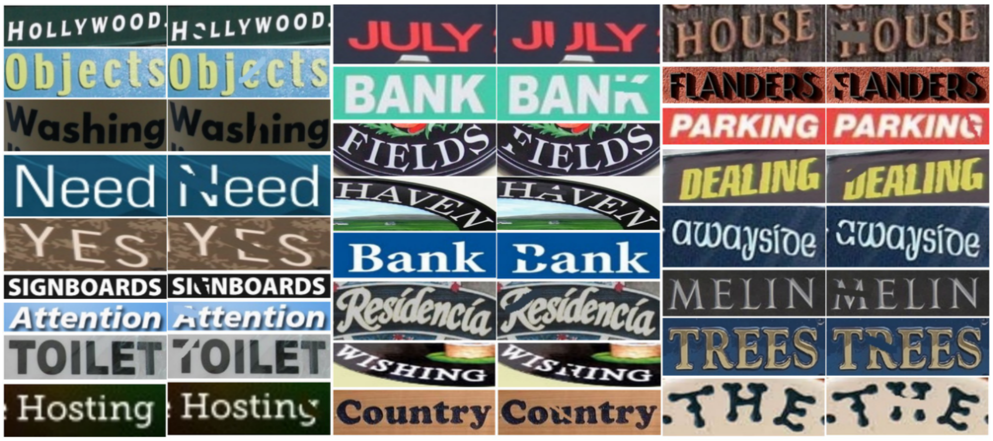

Figure 2: The word images occluded in weak degree.

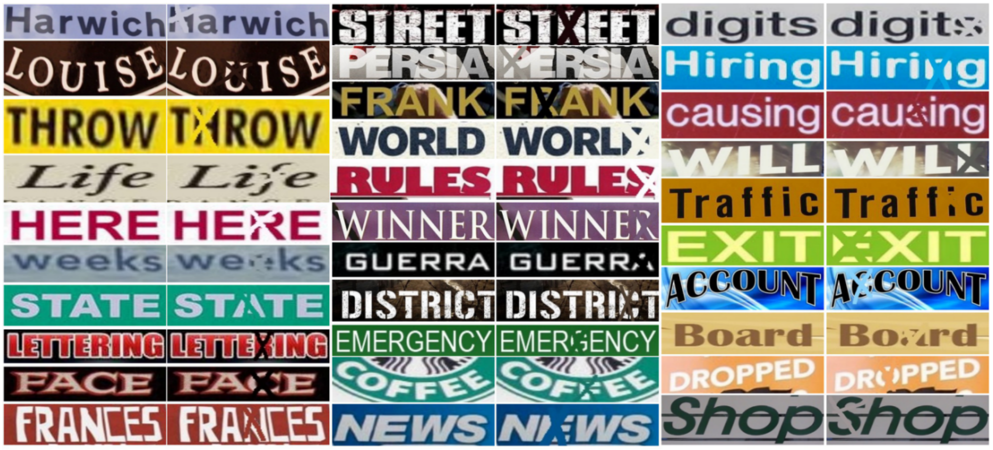

Figure 3: The word images occluded in heavy degree.

3. The Effectiveness on Language-free Dataset

By considering the visual and linguistic information as a union, VisionLAN can adaptively consider less linguistic information for the recognition of language-free text images (*e.g.* numbers, URL, etc.). As the relationship between characters is also helpful, the exploration of linguistic information in the visual context will give low confidence to

Table 3: The results on MLT2019 dataset.

Methods	Accuracy
Baseline + Transformer[3]	72.9
VisionLAN	**74.2**

the Latin character in some time steps for number recognition. We further test the performance of VisionLAN on our own collected ID card dataset, which contains around 31 million cropped images for training and 5000 cropped images for testing. Word-level accuracy is used for evaluation. Compared with RNN-based linguistic learning structure [2], VisionLAN obtains 2.2% improvement (shown in Tab. 2), proving the effectiveness of our approach.

4. The Generalization Ability on MLT2019

To evaluate the performance of VisionLAN on multilingual text recognition, we crop word instances from the images in Task 4 [1] for training and testing. Specifically, we randomly select 1900 images for testing and others are used for training. Words with $length > 25$ are ignored in both training and testing. The number of character categories is calculated from all labels. Word-level accuracy is used for evaluation. In terms of approaching speed and parameters, we implement one transformer unit in GSRM of [3] for fair comparison. As shown in Tab. 3, VisionLAN achieves 1.3 % improvement, which effectively demonstrates the effectiveness of our method.

References

[1] Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowdhury, Dimosthenis Karatzas, Wafa Khlif, Jiri Matas, Umapada Pal, Jean-Christophe Burie, Cheng-lin Liu, et al. Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1582–1587. IEEE, 2019.

[2] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2035–2048, 2018.

[3] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Towards accurate scene text recognition with semantic reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12113–12122, 2020.