Harnessing the Power of Multi-Lingual Datasets for Pre-training: Enhancing Text Spotting Performance: Supplementary Material

Alloy Das [1] Sanket Biswas[2] Ayan Banerjee[2] Josep Lladós[2] Umapada Pal[1] Saumik Bhattacharya[3]

[1]CVPR Unit, Indian Statistical Institute, Kolkata [2]CVC, Universitat Autònoma de Barcelona [3]ECE, Indian Institute of Technology, Kharagpur

1. Some More Insights

Representation quality and robustness. To evaluate the representation quality of the features learned during pre-training and the need for the encoder and dual decoder blocks in the Swin-TESTR pipeline we did an exhaustive evaluation by freezing and unfreezing them. We performed all the possible combinations of frozen and unfrozen decoder and encoder to show the representation quality. Results have been highlighted in Table 1 performed on the Total-Text dataset in the fine-tuning configuration. The consistent performances on both detection and recognition tasks for all the different configurations show the model's robustness.

Table 1. **Evaluating Representations.** Performance obtained by fine-tuning a pre-trained model using frozen backbones. Here "IE", "LD", and "CD" represent Image Encoder, Text Location Decoder, and Text Recognition Decoder respectively also "None" refers to no lexicon has been used.

Blocks			Detection			End-to-End
IE	LD	CD	P	R	F	None
✗	✗	✓	91.67	77.63	84.07	71.88
✗	✓	✗	**93.33**	81.91	87.25	73.74
✗	✓	✓	90.86	84.64	87.64	68.9
✓	✗	✗	91.89	83.51	87.50	74.27
✓	✗	✓	92.52	83.12	87.57	74.56
✓	✓	✗	91.23	84.69	87.84	**75.14**
✓	✓	✓	90.58	**85.46**	**87.95**	74.13

Effectiveness of Swin Transformer backbone over ResNet 50. Connecting remote features with vanilla convolutions can be a challenging task due to their local operation at fixed sizes, such as 3×3. However, in the case of text spotting, it becomes crucial to capture the relationships between different texts. This is because scene text within the same image tends to exhibit similarities in terms of the text background, style, and texture. To address this issue, we have selected a backbone architecture known as Swin-Transformer [9], specifically the Swin-tiny variant, for feature extraction. The Swin-Transformer unit stands out as a small and efficient option that fits our requirements. By incorporating the Swin-tiny unit into our framework, we aim to enhance the ability of our model to understand the contextual connections between texts, thereby improving text spotting performance. In summary, the challenge lies in connecting remote features using conventional convolutions, which have limited receptive fields. However, in the context of text spotting, it is crucial to capture the relationships between different texts. To tackle this, we utilize the Swin-tiny backbone, which is a compact and efficient variant of the Swin-Transformer, to extract features that enable our model better to comprehend the contextual information present in text scenes. we have illustrated how our Swin-tiny backbone manages to generate a better-localized representation over the text than the standard Resnet-50 model. The final output from the last layer of the encoder is then propagated to the next phase. Fig. 1 illustrates how we incorporate two dilated convolution layers, one vanilla convolution layer, and one residual structure into the original Swin-Transformer, which also introduces CNN properties to Transformer.

Figure 1. **Illustration of the effectiveness of the Swin Backbone over Resnet 50.** first one is the original image, the second one is the feature map extracted by Resnet 50. and the last one is the feature map generated by Swin-tiny Backbone

2. Datasets and Experimental Setup

ICDAR 2015 [4] is the official dataset for ICDAR 2015 robust reading competition for regular text spotting. It has 1000 training and 500 testing images. As discussed in [13] it has 6.9 words per image.
Total-Text [1] is a well-known arbitrary-shaped text spotting benchmark with 1255 training and 300 testing images. The text instances in this benchmark are at the word level. As discussed in [13] it has 7.4 words per image.
CTW1500 [7] is another arbitrary-shaped text spotting benchmark consisting of 1000 training and 500 testing images. The text instances in this benchmark are at the multi-word level.
VinText [12] is a recently released Vietnamese text dataset containing arbitrary-shaped text. It consists of 1200 training images and 500 testing images.
Curved SynthText 150K dataset synthesized in [6], consisting of 94723 images with mostly straight text and 54327 images with major curved text.
ICDAR 2017 MLT [11] is a multi-lingual text dataset. It contains 7200 training images and 1800 validation images. We only select the images containing English texts for training. The text instances in this benchmark are at the word level. As discussed in [13] it has 9.5 words per image which is very high.
ReCTS [15] is a Chinese arbitrary-shaped text dataset with 20000 training images and 5000 testing images.

Experimental Setup: One RTX 3080 Ti GPU with a batch size of 1 for an overall 16 days for pre-training phase. For supervised training on the corresponding evaluation datasets to understand where we stand in the competition, we fine-tune Total-Text, and ICDAR2015 with 200K iterations. For the CTW1500 dataset, we fine-tune Swin-TESTR over 2000K iterations with a maximum length of text 100. As it is annotated sentence level it needs more iterations to fine-tune. For Vintext we fine-tune our model with 1000K iterations.

3. More Performance Evaluation and Analysis

Performance Analysis on the Out-of-Vocabulary dataset. Here we have also evaluated the domain adaptation setting in the Out-of-Vocabulary (OOV) text recognition dataset [2]. An OOV dataset in the context of natural language processing and machine learning typically refers to *words that are not in the model's training dictionary* or vocabulary. This situation often arises when the model is used in real-world applications, where it encounters words it has not seen during training. As shown in Table 2, results actually prove again how the MLT17 dataset is beneficial for the domain adaptive pretraining (Synth to Real), justifying our hypothesis that *generally larger and more diverse training sets will result in better OOV performance for text recognition.*

Performance Analysis on the ReCTS dataset In the case of Chinese, where there are thousands of characters and many more possible words, this can be especially important. . The results as shown in Table 3 demonstrate that our model doesn't perform the best compared to the other baselines. We hypothesize it's mainly due to the pre-training with the English Curved SynthText [6]. The performances have the potential to match with the existing SOTA if pre-trained on a more domain-specific Chinese Synthetic dataset. There is more scope for exploration in future works.

4. Qualitative Analysis on Document Layout Analysis

In Figure 2, we present a comprehensive assessment of the reading proficiency achieved by our Swin-TESTR model in the context of document layout analysis. This evaluation revolves around the model's ability to detect various layout regions within documents by leveraging the OCR capabilities embedded in our framework.

Our performance evaluation reveals that Swin-TESTR stands as a robust contender, yielding results that closely rival those obtained through the original document layout analysis OCR. Notably, it surpasses the performance of the original model when it comes to identifying text, mathematical content, and table regions. This achievement underscores Swin-TESTR's prowess in extracting valuable information from documents.

However, it is essential to acknowledge that Swin-TESTR exhibits relatively poorer performance in the identification of separator regions, miscellaneous elements, and image-containing sections. This limitation can be attributed to the model's initialization with text spotting weights, which essentially means it was trained predominantly on text-centric data and, consequently, has limited exposure to instances of these other region types.

References

[1] Chee-Kheng Ch'ng, Chee Seng Chan, and Cheng-Lin Liu. Total-text: toward orientation robustness in scene text detection. *International Journal on Document Analysis and Recognition (IJDAR)*, 23(1):31–52, 2020. 2

[2] Sergi Garcia-Bordils, Andrés Mafla, Ali Furkan Biten, Oren Nuriel, Aviad Aberdam, Shai Mazor, Ron Litman, and Dimosthenis Karatzas. Out-of-vocabulary challenge report. *arXiv preprint arXiv:2209.06717*, 2022. 2

[3] Mingxin Huang, Yuliang Liu, Zhenghao Peng, Chongyu Liu, Dahua Lin, Shenggao Zhu, Nicholas Yuan, Kai Ding, and Lianwen Jin. Swintextspotter: Scene text spotting via better

Table 2. Performance Analysis on Out-of-Vocabulary End-to-End Recognition. Results style: **best**, second best

Method	Recall	Precision	Hmean
Synth-Text	0.0011	0.0033	0.0021
Synth-Text → ICDAR MLT17	0.1146	0.2903	0.1644
Synth-Text → ICDAR15	0.0217	0.1088	0.0361
Synth-Text → ICDAR MLT17 → ICDAR15	0.0563	0.2076	0.0886
Synth-Text → ICDAR MLT17 → Total-Text	0.0446	0.1638	0.0701
Synth-Text → ICDAR MLT17 → CTW1500	0.0399	0.2518	0.0688
Synth-Text → Total-Text	0.0445	0.1779	0.0711
Synth-Text → CTW1500	0.0375	0.1510	0.0601
Mix pre-train	0.0865	0.4674	0.1459
ICDAR MLT17	0.0364	0.1309	0.0433
ICDAR MLT17 → Synth-Text	0.0439	0.1712	0.0699
Synth-Text → ReCTS	0.0848	0.2473	0.1263
Mix pre-train → Fine-tune	**0.2492**	**0.2724**	**0.2603**

Table 3. **Text spotting results on ReCTS**. Results style: **best**, second best

Methods	Detection			1-NED
	P	R	F	
FOTS [5]	78.3	82.5	80.31	50.8
Mask TextSpotter [10]	89.3	88.8	89.0	67.8
AE TextSpotter [14]	92.6	**91.0**	**91.8**	71.8
ABCNet v2 [8]	93.6	87.5	90.4	62.7
Swintextspotter [3]	**94.1**	87.1	90.4	**72.5**
Swin-TESTR	92.61	67.72	78.21	68.12

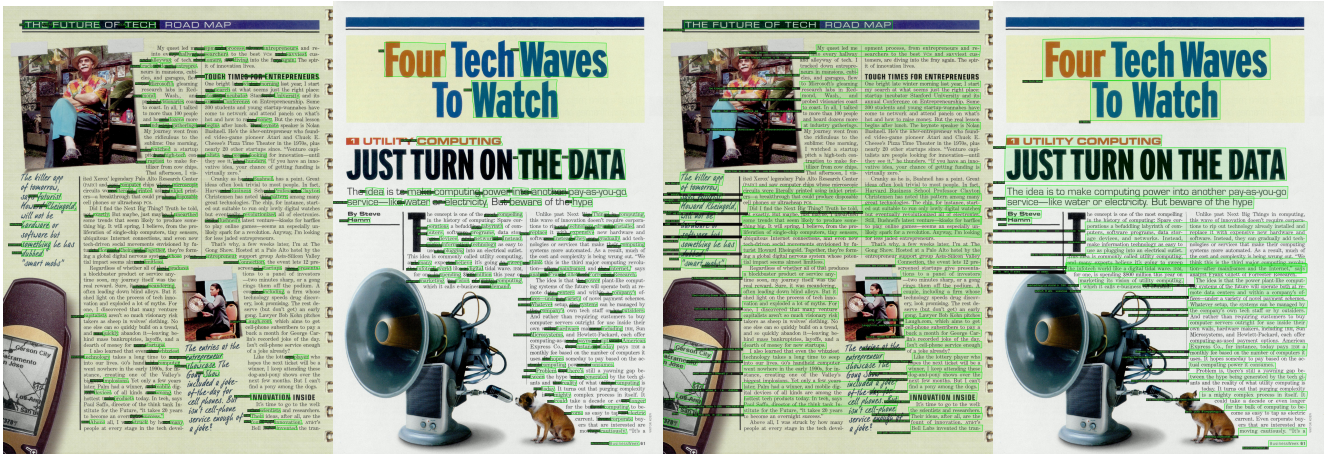
synergy between text detection and text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4593–4603, 2022. 3

[4] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 2

[5] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5676–5685, 2018. 3

[6] Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9809–9818, 2020. 2

[7] Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and



(a) Industrial OCR 1 (b) Industrial OCR 2 (c) Swin-TESTR 1 (d) Swin-TESTR 2

Figure 2. Comparative Analysis of Swin-TESTR as OCR for document layout analysis.

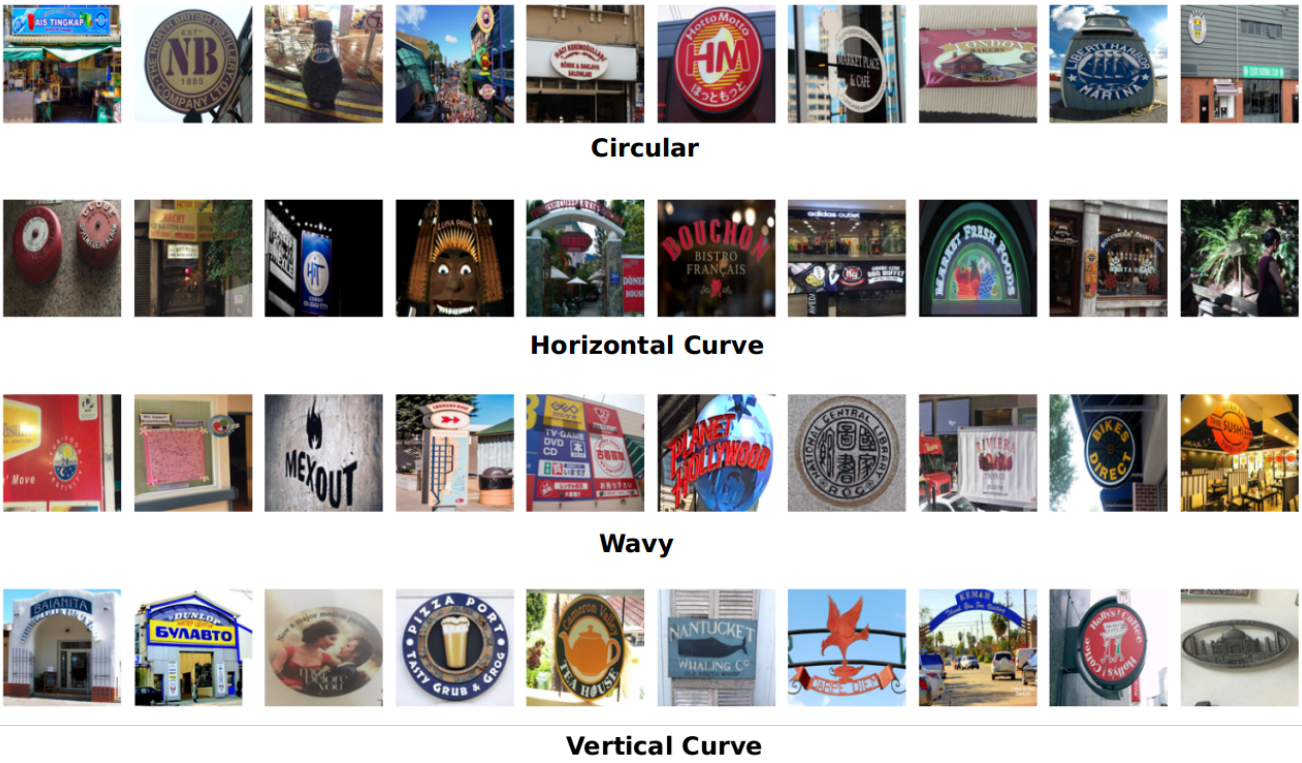


Figure 3. Clusterization of Total-Text through the shape of the text.

Figure 4. Clusterization of CTW1500 through the shape of the text.

(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 5. **Performance Analysis on CTW1500.** Some qualitative test cases of our method on images from CTW1500.

Sheng Zhang. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognition*, 90:337–345, 2019. 2

[8] Yuliang Liu, Chunhua Shen, Lianwen Jin, Tong He, Peng Chen, Chongyu Liu, and Hao Chen. Abcnet v2: Adaptive bezier-curve network for real-time end-to-end text spotting. *arXiv preprint arXiv:2105.03620*, 2021. 3

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

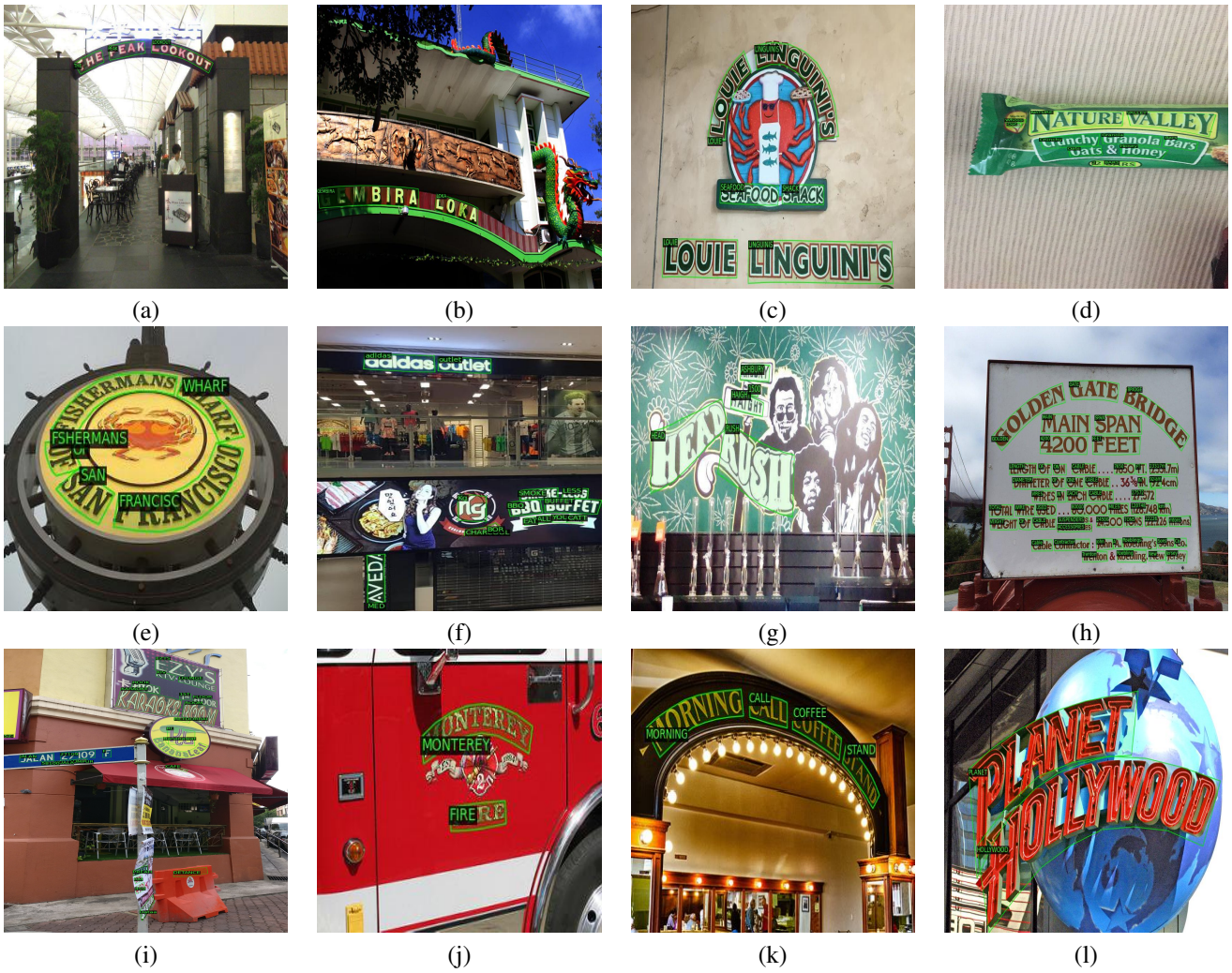
[10] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67–83, 2018. 3

[11] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 1454–1459. IEEE, 2017. 2

[12] Nguyen Nguyen, Thu Nguyen, Vinh Tran, Minh-Triet Tran, Thanh Duc Ngo, Thien Huu Nguyen, and Minh Hoai. Dictionary-guided scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7383–7392, 2021. 2

[13] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8802–8812, 2021. 2

[14] Wenhai Wang, Xuebo Liu, Xiaozhong Ji, Enze Xie, Ding Liang, ZhiBo Yang, Tong Lu, Chunhua Shen, and Ping Luo. Ae textspotter: Learning visual and linguistic representation for ambiguous text spotting. In *Computer Vision–*



(a) (b) (c) (d)

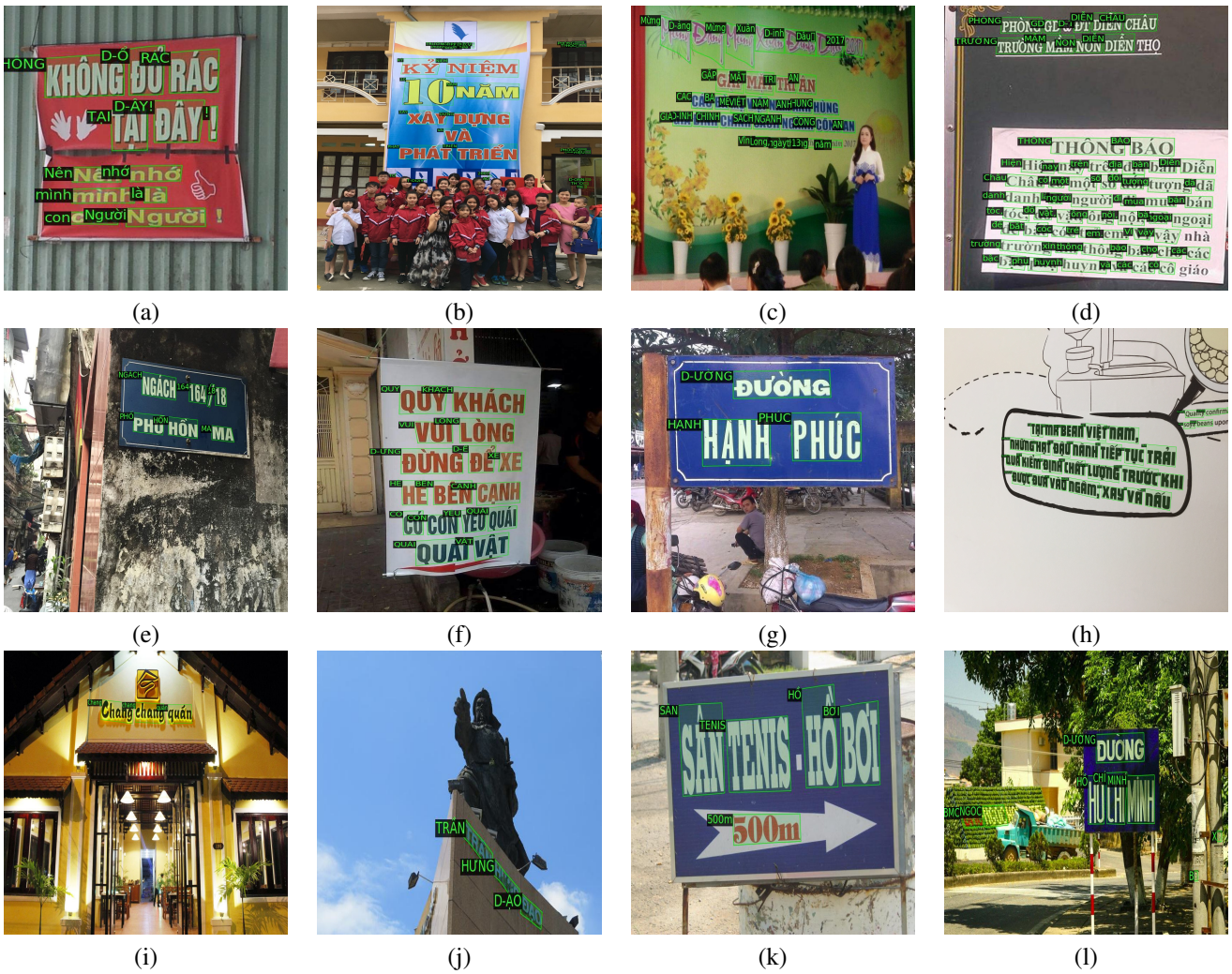
(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 6. **Performance Analysis on TotalText.** Some qualitative test cases of our method on images from TotalText.

ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16, pages 457–473. Springer, 2020. 3

[15] Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 1577–1581. IEEE, 2019. 2



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 7. **Performance Analysis on Low Resource VinText.** Some qualitative test cases of our method on images from the Vietnamese Text dataset.

Figure 8. **Performance Analysis on ICDAR15.** Some qualitative test cases of our method on images from ICDAR15.

(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 9. **Failure Cases.** Illustration of some failure cases of our method a and e from Total Text, b and f are from ICDAR15, c and g are from Vintext and d and h from CTW1500 dataset.