BTS: A Bi-lingual Benchmark for Text Segmentation in the Wild (Supplementary Material)

A. BTS Domain Studies

To further explore the significance of the BTS for text segmentation development, we conduct experiments on HRNetV2-W48 [4] and HRNetV2-W48+OCR [6]. Taking the size of the dataset into consideration, we mainly perform domain studies on four large-scale datasets including TextSeg [5], COCO_TS [1], MLT_S [2], and Total-Text [3].

For TextSeg, we compare the models trained on TextSeg, with models firstly pretrained on BTS, then finetuned on TextSeg, whose results are shown in Tab. 1. For the other three datasets, we compare the models trained on its own training set, with models firstly pretrained on TextSeg and BTS, then finetuned on its own training set, whose results are shown in Tab. 3. The pretrain strategy can not only boost the performance, but also make the model converge faster in the finetune stage.

As Tab. 1 shows, with the help the BTS, the F-score of HRNetV2-W48 is raised from 0.914 to 0.921, and the F-score of HRNetV2-W48+OCR is raised from 0.918 to 0.923. The results of the simple model HRNetV2-W48+OCR with BTS are already comparable to those of the complex model TexRNet+HRNetV2-W48 without BTS. As shown in Tab. 3, on the other three datasets, the pretrained models on TextSeg and BTS are also helpful to improve the fgIoU and the F-score for HRNetV2-W48 and HRNetV2-W48+OCR. Further, on COCO_TS, the fgIoU of the simple model HRNetV2-W48 with TextSeg and BTS is already higher than that of TexRNet+HRNetV2-W48 without TextSeg and BTS, which is a complex model with more parameters. The results of the domain studies demonstrate that the large-scale and high-quality dataset can improve the performance of various models on different datasets, and help the simple model to achieve comparable performance with the complex model without pretraining.

B. The Generalization Ability Studies for PGTSNet

In this section, we aim at evaluating the generalization ability of the proposed PGTSNet. Firstly, we conduct more experiments to show quantitative results of PGTSNet compared with a variety of state-of-the-art methods on Total-Text. Then, we show some qualitative results when applying PGTSNet to unseen scenes which are not included in the training set during the training process.

Table 1. Comparison experiments of training the same model with different datasets on TextSeg.

Method	Train Dataset	TextSeg	
		fgIoU	F-score
HRNetV2-W48	Original	0.850	0.914
	+ BTS	**0.852**	**0.921**
HRNetV2-W48+OCR	Original	0.860	0.918
	+ BTS	**0.860**	**0.923**
TexRNet+HRNetV2-W48	Original	0.868	0.924

Table 2. The comparison experiments of PGTSNet with the state-of-the-art methods on Total-Text.

Method	Train Dataset	fgIoU	F-score
PSPNet	Original	-	0.740
SMANet	Original	-	0.770
DeeplabV3+	Original	0.744	0.824
HRNetV2-W48	Original	0.753	0.825
HRNetV2-W48	+TextSeg+BTS	0.774	0.840
HRNetV2-W48+OCR	Original	0.762	0.832
HRNetV2-W48+OCR	+TextSeg+BTS	0.788	0.847
TexRNet	Original	0.765	0.844
TexRNet	+TextSeg+BTS	0.774	0.846
PGTSNet	+TextSeg+BTS	**0.791**	**0.847**

B.1. Comparisons on Total-Text

Taking the annotation quality of the dataset into consideration, we conduct experiments on Total-Text. As the dataset only contains the word-level annotations, we remove the character classifier for TexRNet and PGTSNet in the experiments, and utilize DeeplabV3+ as the backbone of all these methods for fair comparisons. The results of PSPNet and SMANet are from [1] and [2], respectively, which are trained on ICDAR13 FST and Total-Text augmented with SynthText. As Tab. 2 shows, the fgIoU and F-score of PGTSNet are both higher than those of the compared methods on Total-Text, demonstrating the effectiveness and robustness of the proposed PGTSNet.

Table 3. Comparison experiments of training the same model with different datasets on COCO_TS, MLT_S, and Total-Text, respectively.

Method	Train Dataset	COCO_TS		MLT_S		Total-Text	
		fgIoU	F-score	fgIoU	F-score	fgIoU	F-score
HRNetV2-W48	Original	0.689	0.629	0.832	0.836	0.753	0.825
	+ TextSeg +BTS	**0.742**	**0.680**	**0.845**	**0.839**	**0.774**	**0.840**
HRNetV2-W48+OCR	Original	0.695	0.627	0.834	0.838	0.762	0.832
	+ TextSeg +BTS	**0.762**	**0.669**	**0.856**	**0.841**	**0.788**	**0.847**
TexRNet+HRNetV2-W48	Original	0.724	0.720	0.861	0.865	0.785	0.848

Figure 1. One failure case for our proposed method. From top to down are the original image, the result of TexRNet (trained with BTS), and the result of PGTSNet, respectively.

Figure 2. An example of pixel-level mask annotating.

B.2. Comparison on New Scene

We also apply PGTSNet to unseen scenes which are not included in the training set of either TextSeg or BTS, such as the film frames with subtitles. The qualitative comparisons are shown in Fig. 3. From Fig. 3 we can see that the proposed PGTSNet can extract the precise segmentation masks for various and even tiny characters of the images from unseen scenes, demonstrating the generalization ability of the proposed PGTSNet.

B.3. The failure cases, limitations, and social impact

Fig. 1 shows a failure case, where our proposed model PGTSNet fails to handle texts with some artistic effects such as a border of different color. However, PGTSNet still performs better than the compared method TexRNet on these cases. All the images in BTS are manually annotated by humans in three levels, including the pixel-level, the character-level, and the line-level annotations. There is no automatic algorithms or out-of-the-box models involved during the labeling process. PhotoShop is the main tool. As shown in Fig. 2, the pencil tool in Photoshop is utilized to assist the annotators to label pixel-level mask annotations for texts. The labeling workflow ensures all annotations to be made in relatively high quality and the benchmark to be highly-reliable. In BTS, about 70% images were captured by the annotators using the cameras of their cellphones; the left 30% images were collected from several websites without copyright constraints. We confirm that the copyright issue in BTS is eliminated. And it does not have any sensitive information including personal privacy, political issues, vulgar and violent contents, etc. Thus, it can be safely used for academic research.

C. Future dataset

We will explore more scenes such as graffiti walls and posters which contain more designed texts, and also add more character classes and try to make the classes more balanced in the future work.

References

[1] S. Bonechi, P. Andreini, M. Bianchini, and F. Scarselli. *COCO-TS Dataset: Pixel–Level Annotations Based on Weak Supervision for Scene Text Segmentation.* Artificial Neural Networks and Machine Learning – ICANN 2019: Image Processing, 2019. 1

[2] S. Bonechi, M. Bianchini, F. Scarselli, and P. Andreini. Weak supervision for generating pixel–level annotations in scene text segmentation. *Pattern Recognition Letters*, 138, 2020. 1

[3] C. K. Ch'Ng and C. S. Chan. Total-text: A comprehensive dataset for scene text detection and recognition. *IEEE*, 2018. 1

Figure 3. Qualitative comparisons between PGTSNet and TexRNet. From top to bottom, row-1,4,7 are input images, row-2,5,8 are predicted masks of TexRNet trained on BTS and TextSeg, row-3,6,9 are predicted masks of PGTSNet trained on the same datasets.

[4] J. Wang, K. Sun, T. Cheng, B. Jiang, and B. Xiao. Deep high-resolution representation learning for visual recognition. 2019. 1

[5] X. Xu, Z. Zhang, Z. Wang, B. Price, and H. Shi. Rethinking text segmentation: A novel dataset and a text-specific refinement approach. 2020. 1

[6] Y. Yuan, X. Chen, and J. Wang. Object-contextual representations for semantic segmentation. 2019. 1