Rethinking Text Segmentation: A Novel Dataset and A Text-Specific Refinement Approach - Supplementary Material -

Xingqian Xu[1], Zhifei Zhang[2], Zhaowen Wang[2], Brian Price[2], Zhonghao Wang[1], Humphrey Shi[3,1]

[1]UIUC, [2]Adobe Research, [3]University of Oregon

A. Cosine Similarity vs. Accuracy

Recall that in our paper, we claim that the cosine similarity of the predicted mask is inversely correlated with its accuracy. To provide solid evidence on our claim, we produce the following experiment in which we compute the cosine-similarity, *i.e.* CosSim(x'_{sem}), and the fgIoU score on each image and plot their relation in Figure 1. According to our plot, there is a clear downward trend between Cosine Similarity and fgIoU, from which our claim can be verified.

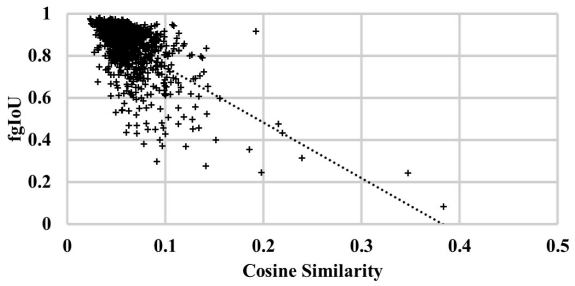


Figure 1: The relation between cosine similarity and fgIoU from predicted masks. In our case, the cosine similarity is computed from the vectorized text prediction and vectorized background prediction.

B. TextSeg Domain Studies

To further show that our TextSeg is a strong complementary towards many prior datasets, we performed domain studies in the same fashion as [2]. The goals of our experiments are two-fold, a) to compare fairly with SMANet [2] on ICDAR13 FST [4] and Total-Text [1] under the same dataset setting, and b) to show the performance boost by including our dataset TextSeg in the training process. The experiments were carried out using our proposed TexRNet with nearly the same experiment settings as explained in Session 5.1 of the main paper. The only differences are: we disabled the discriminator loss (*i.e.* $\mathcal{L}_{dis}$) and we used 20,500 iterations in the ICDAR13 FST experiments. Such changes were due to the fact that no character bounding polygons were provided in COCO_TS [1], MLT_S [2], and more images could be used to avoid overfits. As Table 1 and Table 2 shows, our TexRNet reached F-score 0.866 on ICDAR13 FST and 0.844 on Total-Text, exceeding SMANet using the same dataset combination in training. Meanwhile, we demonstrated an extra 3.20% and 2.93% increase in fgIoU when included our TextSeg in training on ICDAR13 FST and Total-Text, correspondingly.

C. Visual Comparison on TexRNet

To help to understand the key structure of our TexRNet, we extract activation maps from intermediate stages of TexRNet to show how the low confidence text regions in the initial prediction are re-activated using our key pooling and attention module.

Figure 2: Segmentation samples in which the key pooling and attention mechanism in TexRNet helps re-activate low-confidence text regions and achieves better performance. From left to right are original images, ground truth labels, initial predictions, gradients of activation, and refined predictions. In the gradients of activation plot (*i.e.* the 4^{th} column), red indicates positive score changes and blue indicates negative score changes.

Method	Train Dataset	fgIoU	F-score
SMANet [2]	ICDAR13 FST	-	0.713
	ICDAR13 FST + Synth [3]	-	0.785
	ICDAR13 FST + COCO_TS + MLT_S	-	0.858
TexRNet (Ours)	ICDAR13 FST	73.38	0.850
	ICDAR13 FST + COCO_TS + MLT_S	76.68	0.866
	ICDAR13 FST + COCO_TS + MLT_S + TextSeg (ours)	78.65	0.871
	ICDAR13 FST + TextSeg (Ours)	**79.88**	**0.887**

Table 1: Domain studies on ICDAR13 FST in which models are training with different datasets and are evaluated on ICDAR13 FST test set.

Method	Train Dataset	fgIoU	F-score
SMANet [2]	Total-Text	-	0.741
	Total-Text + Synth [3]	-	0.770
	Total-Text + COCO_TS + MLT_S	-	0.781
TexRNet (Ours)	Total-Text	78.47	0.848
	Total-Text + COCO_TS + MLT_S	77.40	0.844
	Total-Text + COCO_TS + MLT_S + TextSeg (Ours)	80.01	**0.858**
	Total-Text + TextSeg (Ours)	**80.33**	0.856

Table 2: Domain studies on Total-Text in which models are training with different datasets and are evaluated on Total-Text test set.

In particular, the 4^{th} column of Figure 2 highlights the gradients of activation scores between the initial predictions (*i.e.* x'_{sem}) and the activation maps prior to the concatenation and the refinement layers (*i.e.* x_{att}).

D. Visualization on Text Removal

This session shows extra samples from our text removal experiment. Recall that we predicted text masks from our TexRNet and used them as inputs for Deep Image Prior [5]. The TexRNet was trained on TextSeg train and validation sets, while all demo images were from TextSeg test set. We also produced examples using ground truth bounding polygons as alternative inputs. As shown in Figure 3, text-free images generated using our predicted mask has the best performance.

E. Visualization on Text Style Transfer

This session shows extra style transfer samples using predicted text masks from our TexRNet and text style transfer network Shape-Matching GAN [6]. Same as text removal, our model was trained on TextSeg train and validation sets, and predicted on the test set. For each sample, the original image, the predicted text mask, and the final result are shown from left to right. We show three styles in total, which is fire (Figure 4a), maple (Figure 4b) and water (Figure 4b).

References

[1] Simone Bonechi, Paolo Andreini, Monica Bianchini, and Franco Scarselli. Coco_ts dataset: Pixel–level annotations based on weak supervision for scene text segmentation. In *International Conference on Artificial Neural Networks*, pages 238–250. Springer, 2019.

[2] Simone Bonechi, Monica Bianchini, Franco Scarselli, and Paolo Andreini. Weak supervision for generating pixel–level annotations in scene text segmentation. *Pattern Recognition Letters*, 138:1–7, 2020.

[3] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2315–2324, 2016.

[4] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1484–1493. IEEE, 2013.

[5] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454, 2018.

[6] Shuai Yang, Zhangyang Wang, Zhaowen Wang, Ning Xu, Jiaying Liu, and Zongming Guo. Controllable artistic text style transfer via shape-matching gan. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4442–4451, 2019.

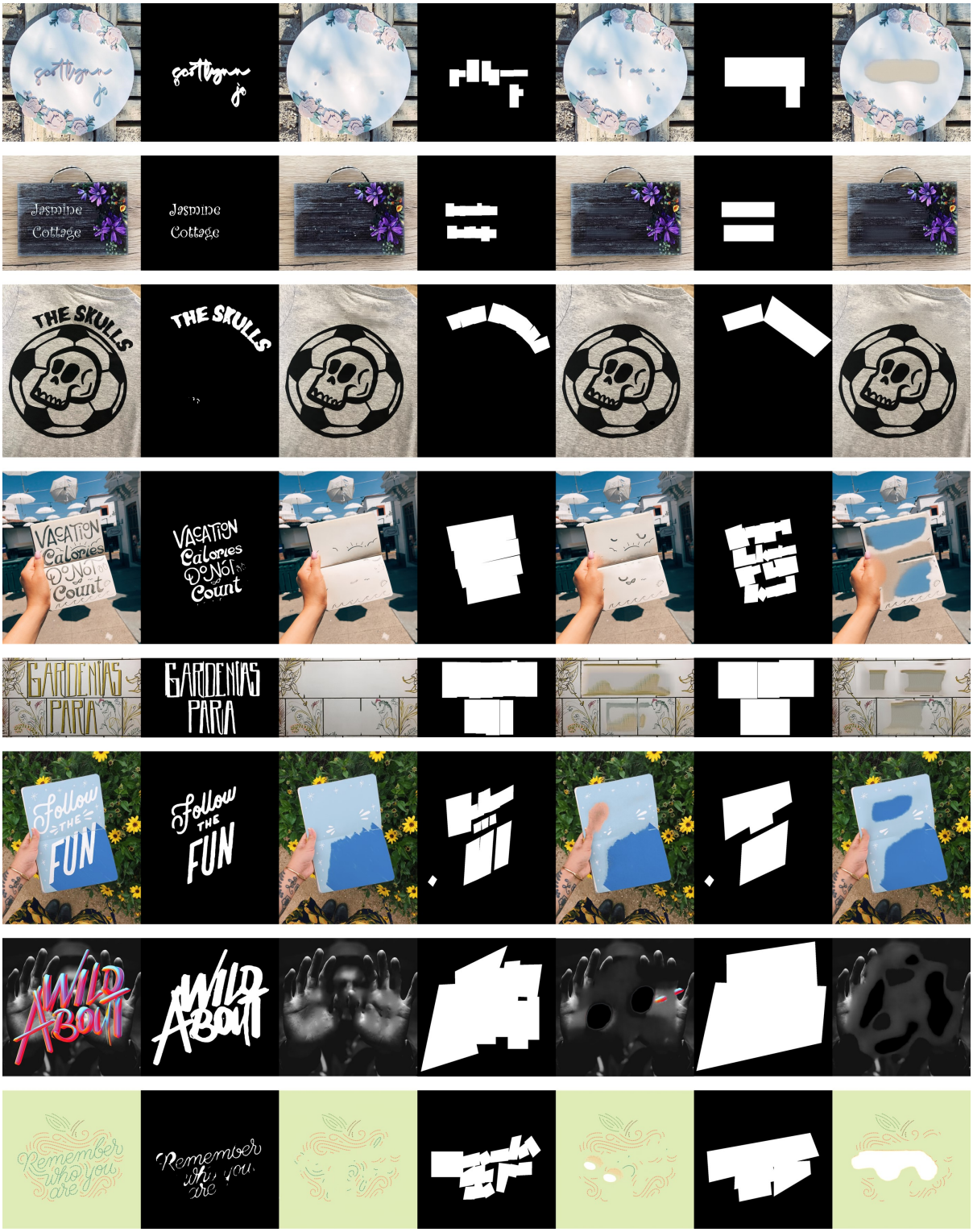
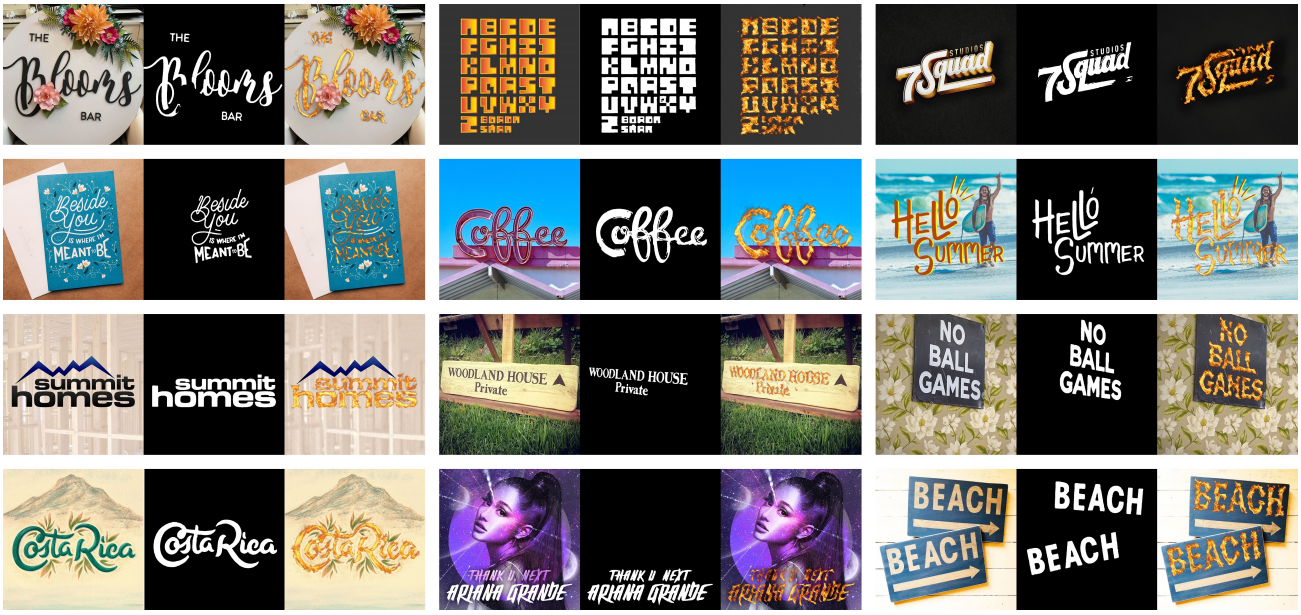
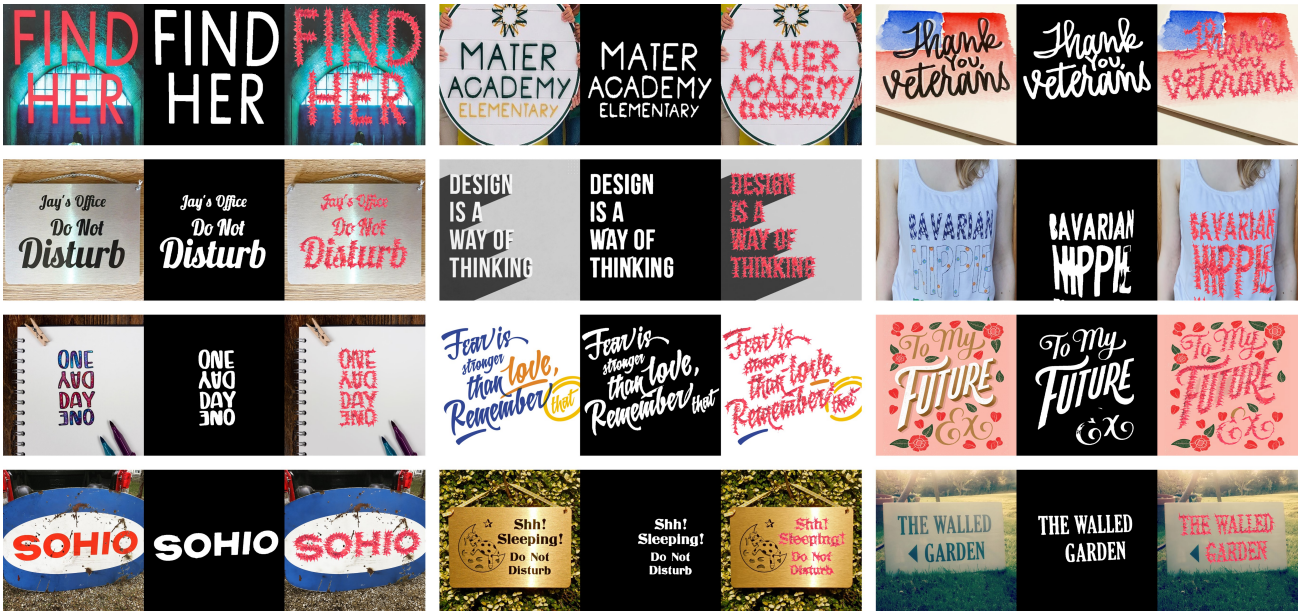


Figure 3: Text removal visualization using predicted text masks from our TexRNet and inpainting network Deep Image Prior [5]. For each sample, the left-to-right ordering of the plots are as such: original image; predicted mask; text removal using mask; ground truth character bounding polygon (char-bpoly); text removal using char-bpoly; ground truth word bounding polygon (word-bpoly); text removal using word-bpoly.

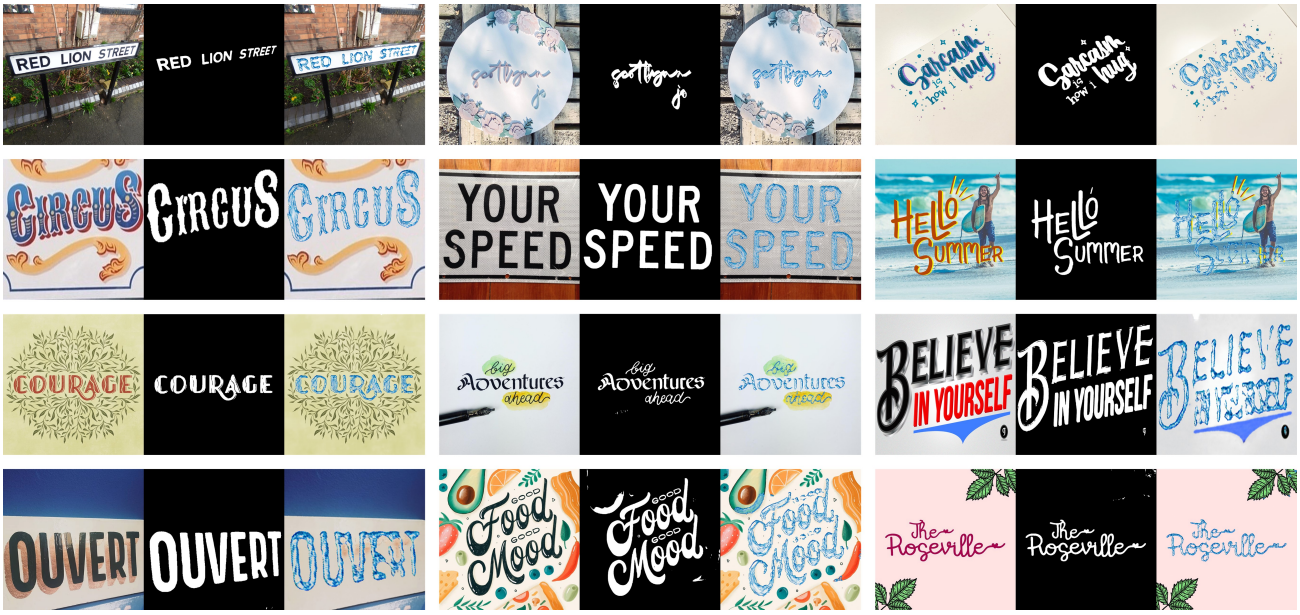


(a) Fire style



(b) Maple style

Figure 4: Style transfer visualization using predicted text masks from our TexRNet and text style transfer network Shape-Matching GAN [6]. For each sample, the original image, the predicted text mask, and the final result are shown from left to right.



(c) Water style

Figure 4: Style transfer visualization using predicted text masks from our TexRNet and text style transfer network Shape-Matching GAN [6]. For each sample, the original image, the predicted text mask, and the final result are shown from left to right.