Supplemental Material: WeCromCL: Weakly Supervised Cross-Modality Contrastive Learning for Transcription-only Supervised Text Spotting

Jingjing Wu*[1], Zhengyao Fang*[1], Pengyuan Lyu*[2], Chengquan Zhang[2], Fanglin Chen[1], Guangming Lu[1], and Wenjie Pei†[1]

[1] Harbin Institute of Technology , Shenzhen, China
[2] Department of Computer Vision Technology, Baidu Inc.
{jingjingwu_hit, zhengyaonineve, wenjiecoder}@outlook.com, lvpyuan@gmail.com, zhangchengquan@baidu.com, {chenfanglin, luguangm}@hit.edu.cn,

1 Instantiation by Adapting SRSTS

SRSTS-A consists of Image Encoder, Anchor Estimator, Sampling Module and Recognition Module, which is consistent with SRSTS. Taking extracted features from Image Encoder, Anchor Estimator predicts the anchor point for each text transcription. Meanwhile, Sampling Module performs sampling around each anchor and provides Recognition Module with the sampled features for text decoding. While the modeling details of SRSTS can be found in the corresponding paper, we elaborate on the differences between it and the adapted SRSTS-A.

We adapt SRSTS to SRSTS-A with three major modifications. First, the detection branch of SRSTS using text boundaries is dropped in our text spotter. Second, we incorporate several practical techniques of DeepSolo (ResNet-50) [5] into SRSTS-A for system enhancement, including the image encoder, data augmentation and optimizating strategies. Third, we modify the loss function for Anchor Estimator $\mathcal{L}_c$ of SRSTS from Dice loss [3] to Focal loss [2] to adapt to the supervision variation from text boundaries to anchor points for better convergence. To be specific, given a feature map $\mathbf{P}$, Anchor Estimator of SRSTS-A learns a confidence map to indicate the probability of each pixel to be an anchor. It employs a 1×1 convolutional layer followed by Sigmoid function to generate the confidence map $\mathbf{C}$. We use Focal loss to optimize the parameters of Anchor Estimator:

$$\begin{aligned}\mathcal{L}_c = \sum_{i=0}^{w-1}\sum_{j=0}^{h-1} &-\alpha \mathbf{C}_{\text{gt}}(i,j)(1-\mathbf{C}(i,j))^{\gamma}\log(\mathbf{C}(i,j)) \\ &-(1-\alpha)(1-\mathbf{C}_{\text{gt}}(i,j))\mathbf{C}(i,j)^{\gamma}\log(1-\mathbf{C}(i,j)),\end{aligned} \tag{1}$$

* Authors contribute equally.
† Corresponding author.

where α and γ are weighting factors for focal loss. $\mathbf{C}_{\text{gt}}$ is pseudo groundtruth for the confidence map constructed from the obtained anchor point by *WeCromCL*: the anchor point is assigned 1 and other pixels are assigned 0.

Integrating *WeCromCL* and SRSTS-A, we obtain the optimized system for transcription-only supervised text spotting.

2 Implementation Details

Implementation details of *WeCromCL*. We firstly pre-train *WeCromCL* on synthetic datasets (Synthtext [1] and Curved Synthetic Dataset) for 200,000 steps with batch size set to be 16. The input size is set to be (640, 640) for fast convergence. Then it will be fine-tuned on the training set of each dataset for 80,000 steps respectively with batch size set to be 4. The following data augmentation strategies are conducted during training: 1) randomly resize the short side of the input image to a range from 640 to 896 while keeping the longer side shorter than 1,280; 2) randomly rotate the input image; 3) randomly apply blur and color jitter. Our method is optimized by SGD with initial learning rate 1e-3 on synthetic datasets and 1e-4 on specific real word dataset. The weight decay is set to be 0.0001 and momentum is set to be 0.9. The learning rate is delayed with a 'poly' strategy. When inferring images to obtain the pseudo location labels, we resize the longer side of input image for ICDAR 2013, ICDAR 2015 to 1152 and 1696, and the shorter side of input image for Total-Text and CTW1500 to 896 and 992.

Implementation details of spotting. In the text spotting stage, our text spotter is supervised by the obtained pseudo location labels. The text spotter is pre-trained on the joint training dataset that contains Curved Synthetic Dataset, ICDAR 2017 MLT, ICDAR 2013, ICDAR 2015, and Total-Text with pseudo location labels for 425,000 steps at first. For word-level benchmarks, our text spotter is fine-tuned on the training set of specific benchmark for 3,000 steps. For CTW 1500, we use line-level text transcriptions of SynthText [1] to generate line-level pseudo location labels and further train the text spotter for 100,000 steps based on the obtained pseudo line-level location labels. Finally, the pre-trained model is further fine-tuned on CTW 1500 training set for 20,000 steps. Adam is used as optimizer. The learning rate is set the same as Deepsolo, and the same data augmentation is used except for Random Cropping operation being removed. In the testing phrase, we resize the shorter side of input image to 864, 864, 1440 and 576 for ICDAR 2013, ICDAR 2015, Total-Text and CTW1500 respectively.

3 Ablation Studies of WeCromCL

Enhancing Fully-Supervised Spotting. Our *WeCromCL* can efficiently generate pseudo location labels from text-image pairs with no annotation cost. Thus we can use it for pseudo data generation and investigate whether it can improve

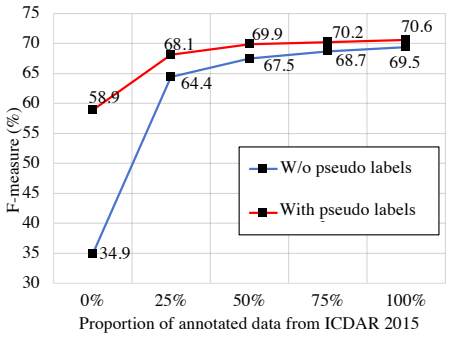

Fig. 1: Effectiveness of our proposed *WeCromCL* on full supervised spotting method. We pre-train SRSTS v2 on Curved Synthetic Dataset and fine-tune it on varying proportion of ICDAR 2015 gt and fixed amount of pseudo labeled data.

the performance of fully supervised single-point spotter. Specifically, we generate pseudo labels for MLT, ICDAR 2013, Total-Text, and TextOCR [4]. Then we reproduce SRSTS v2 based on the Deepsolo framework [5] and pre-train it with Curved Synthetic Dataset. Note that SRSTS v2 can perform text recognition relying only on single point. During the fine-tuning stage, we train it with increasing annotated data from ICDAR 2015 and meanwhile evaluate the effect of adding fixed amount of (sufficient) pseudo-labeled data generated by *WeCromCL*. Figure 1 shows that *WeCromCL* can indeed improve the recognition performance of SRSTS v2, especially when the annotated data is not insufficient. Another interesting observation is that the performance of SRSTS v2 is quite limited when trained only on the synthetic data due to large data distribution gap between synthetic and real-world data. However, its performance is significantly improved when fine-tuned on the real-world data pseudo-labeled with our *WeCromCL*, while no human annotation cost is introduced.

Comparison with Other Keypoint Localization / Pseudo-labeling Methods. To the best of our knowledge, our approach is the first to generate localization pseudo labels using text-only supervision. Our approach's efficacy is evidenced by comparing our pseudo label generation method with those derived from the attention map of another weak supervision method, oCLIP, as presented in Table 3 of our submission. Here, we further include another weak supervision method, NPTS, and evaluate the impact of pseudo labels generated by different methods on the final spotting performance. The experimental results are presented in Table 2 and rows 1, 2, and 3 of Table 1, confirming the effectiveness of our method.

Impact of Anchor Quality on Spotting Results. In Table 1, we compare our method's accuracy (row 3) to ground truth anchor points (row 7) and perturbed anchor points (rows 8, 9) in the spotting task. With α representing the perturbation degree, where α is set to 0.3, the offset from ground truth follows a Gaussian distribution within 0.3 times the text box size. Results show our method's accu-

Table 1: Spotting results under different pseudo-label generation methods and ablation conditions on ICDAR 2015. 'gt' and 'pse' indicates actual ground truth and pseudo labels, respectively. Subscripts 'trans', 'point', and 'box' specify text-only, single-point, and bounding box annotations, respectively. Notably, '$\text{pse}_{\text{point}}$' utilizes the same text-only annotation information as 'gt_{trans}'.

Row	Method	Data		α	S	W	G
		Synth	Real				
1	oCLIP	$\text{pse}_{\text{point}}$-150k	$\text{pse}_{\text{point}}$-11k	–	60.9	57.7	51.7
2	NPTS	$\text{pse}_{\text{point}}$-150k	$\text{pse}_{\text{point}}$-11k	–	72.7	68.4	61.7
3	Ours	$\text{pse}_{\text{point}}$-150k	$\text{pse}_{\text{point}}$-11k	–	82.1	76.1	68.8
4	Ours	gt_{point}-150k	$\text{pse}_{\text{point}}$-11k	–	82.8	76.3	69.4
5	Ours	gt_{point}-150k	gt_{point}-11k	–	84.8	78.4	71.4
6	Ours	gt_{box}-150k	gt_{box}-11k	–	**86.8**	**82.4**	**77.8**
7	Ours	$\text{pse}_{\text{point}}$-150k	gt_{point}-11k	0	82.9	77.0	69.8
8	Ours	$\text{pse}_{\text{point}}$-150k	gt_{point}-11k	0.3	79.1	74.6	68.0
9	Ours	$\text{pse}_{\text{point}}$-150k	gt_{point}-11k	1	42.0	39.9	36.2

Table 2: Comparison of Pseudo Label Quality on ICDAR 2015.

Set	Method	P	R	F
Training	oCLIP	51.1	35.4	41.9
	NPTS	47.6	48.2	47.9
	WeCromCL	**91.4**	**86.0**	**88.6**

racy is close to using ground truth (82.1/76.1/68.8 vs 82.9/77.0/69.8), confirming the accuracy of our pseudo labels. In addition, we trained with ground truth from synthetic data and pseudo label from real-word data (row 6 in Table 1, which does result in a slight improvement. However, this slight difference underscores: 1) the high quality of our method's pseudo labels; 2) our method's effectiveness without relying on explicit position labels.

Overall Gap between Fully Supervised Methods. In Table 1, we observe that the performance gap primarily stems from two factors: pseudo label accuracy (rows 3, 4, 5) and the positional supervision method (rows 5, 6). Comparing ground truth point supervision to pseudo labels supervision, the performance gaps are 2.7, 2.3, and 2.6, respectively. Similarly, the gaps between box supervision and point supervision are 2, 4, and 6.4, highlighting the importance of detailed positional information. Moving forward, we aim to explore methods for obtaining high-quality pseudo labels at the box level.

4 Visualization Results.

4.1 Visualization of Activation Maps

To better illustrate the localization performance of *WeCromCL*, we show sufficient activation maps generated by *WeCromCL* in Figure 2. We can observe

Fig. 2: Visualization of activation maps learned by *WeCromCL*. Our *WeCromCL* can handle various complex cases, such as text with artistic fonts, curved text, long text, and small text. Given a text transcription, *WeCromCL* can generate corresponding activation map in which the highly activated region is identified as the anchor point for this transcription.

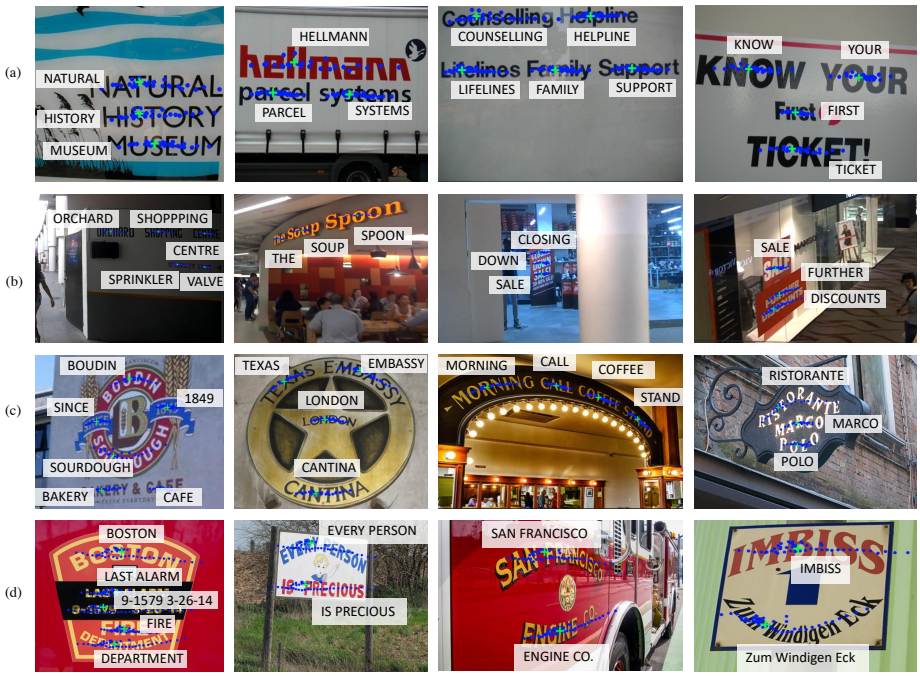

Fig. 3: Visualization of text spotting results on four benchmarks: (a) ICDAR 2013, (b) ICDAR 2015, (c) Total-Text and (d) CTW1500. The green '+' represents the estimated anchor point for each text instance. The blue dots denote the sampled points.

that our *WeCromCL* successfully locates the text region when given a text query. Even when the queried text is small and fuzzy within the image, by enlarging the input image, *WeCromCL* is still capable of successfully locating the most relevant position associated with the queried text. The most activated pixel with the peak value in each activation map is identified as the anchor point for the corresponding transcription. The obtained anchor points are further used as pseudo location labels to supervise the learning of text spotter in the text spotting stage.

4.2 Visualization of Text Spotting Results

Some text spotting results are shown in Figure 3. Our text spotter is learned under the supervision of pseudo location labels obtained by *WeCromCL*. As can be easily seen, the proposed transcription-only supervised text spotter can achieve satisfactory performance even when facing challenging cases such as tiny text, fuzzy text, curved text and long text. With the provided precise pseudo location labels as supervision, our text spotter learns to locate the text instance precisely and successfully performs sampling for text recognition. The visualization of text spotting results intuitively demonstrate the effectiveness and robustness of proposed transcription-only supervised text spotter.

References

1. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: CVPR (2016) 2
2. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017) 1
3. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). IEEE (2016) 1
4. Singh, A., Pang, G., Toh, M., Huang, J., Galuba, W., Hassner, T.: Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In: CVPR (2021) 3
5. Ye, M., Zhang, J., Zhao, S., Liu, J., Liu, T., Du, B., Tao, D.: Deepsolo: Let transformer decoder with explicit points solo for text spotting. In: CVPR (2023) 1, 3