Seq-UPS: Sequential Uncertainty-aware Pseudo-label Selection for Semi-Supervised Text Recognition
Supplemental Material

Gaurav Patel Jan Allebach Qiang Qiu
School of Electrical and Computer Engineering, Purdue University, USA
{gpatel10, allebach, qqiu}@purdue.edu

Layers	Configurations	Output
Input	grayscale image	100×32
Conv1	$c:32$ $k:3\times3$	100×32
Conv2	$c:64$ $k:3\times3$	100×32
Dropout	-	100×32
Pool1	$k:2\times2$ $s:2\times2$	50×16
Block1	$\begin{bmatrix} c:128, k:3\times3 \\ c:128, k:3\times3 \end{bmatrix} \times 1$	50×16
Conv3	$c:128$ $k:3\times3$	50×16
Dropout	-	50×16
Pool2	$k:2\times2$ $s:2\times2$	25×8
Block2	$\begin{bmatrix} c:256, k:3\times3 \\ c:256, k:3\times3 \end{bmatrix} \times 2$	25×8
Conv4	$c:256$ $k:3\times3$	25×8
Dropout	-	25×8
Pool3	$k:2\times2$ $s:1\times2$ $p:1\times0$	26×4
Block3	$\begin{bmatrix} c:512, k:3\times3 \\ c:256, k:3\times3 \end{bmatrix} \times 5$	26×4
Conv5	$c:512$ $k:3\times3$	26×4
Dropout	-	26×4
Block4	$\begin{bmatrix} c:512, k:3\times3 \\ c:512, k:3\times3 \end{bmatrix} \times 3$	26×4
Conv6	$c:512$ $k:2\times2$ $s:1\times2$ $p:1\times0$	27×2
Conv7	$c:512$ $k:2\times2$ $s:1\times2$ $p:0\times0$	26×1
Dropout	-	26×1

Table 1. ResNet architecture configuration for the text recognition model. Here, c, k, s, and p stand for no. of channels, filter size, stride, and padding, respectively.

A. Dataset Descriptions

A.1. Handwriting Recognition Datasets

CVL [13]: 310 individual writers contributed to this handwritten English text dataset, which was divided into two parts: training and testing. 27 of the writers created 7 texts, while the remaining 283 created 5 texts.

IAM [15]: 657 different writers contributed to this English handwritten text dataset, which was partitioned into writer independent training, validation, and test.

A.2. Scene-Text Datasets

ICDAR-15 (IC15) [10]: The images in the dataset were gathered by people wearing Google Glass, therefore many of the images have perspective inscriptions and some are fuzzy. It includes 4,468 training images and 2,077 evaluation images.

ICDAR-13 IC13 [11]: The dataset was created for the ICDAR 2013 Robust Reading competition. It contains 848 images for training and 1,015 images for evaluation.

IIIT5k-Words (IIIT) [16]: Google image searches with query phrases like "billboards" and "movie posters" yielded the text-images. It includes 2,000 training photos and 3,000 evaluation images.

Street View Text (SVT) [25]: The dataset is prepared based on Google Street View and includes text included in street photos. It includes 257 training images and 647 evaluation images.

SVT Perspective (SVTP) [18]: Similar to SVT, SVTP is gathered from Google Street View. In contrast to SVT, SVTP features a large number of perspective texts. It includes 645 images for evaluation.

CUTE80 (CUTE) [19]: CUTE contains curved text images. The images are captured by a digital camera or collected from the Internet. It contains 288 cropped images for evaluation.

COCO-Text (COCO) [24]: This dataset is created from text instances from the original MS-COCO dataset [14].

RCTW [20]: RCTW stands for the **R**eading **C**hinese **T**ext in the **W**ild dataset. Primarily containing Chinese text. Nonetheless, we used the non-Chinese text images in the training set.

Uber-Text (Uber) [28]: Bing Maps Streetside was used to obtain Uber-Text image data. Many of them are house numbers, while others are text on billboards.

Arbritary-shaped Text (ArT) [5]: This dataset contains images with perspective, rotation, or curved text.

Large-scale Street View Text (LSVT) [22, 23]: Data collected from the streets in China. Thus, most of the text is in Chinese.

Multi-Lingual Text (MLT) [17]: This dataset is created to recognize multi-lingual text. It consists of text images from seven languages: Arabic, Latin, Chinese, Japanese, Korean, Bangla, and Hindi.

Reading Chinese Text on Signboard (ReCTS) [27]: Created for the Reading Chinese Text on Signboard competition. It features a large number of irregular texts that are grouped in various layouts or written in different typefaces.

For further extensive details on the used scene-text datasets and the adopted preprocessing we refer the readers to [2, 3].

B. Text-Recongnition Model Architecture

We adopt the best performing recognition model used in [1], [2], and [3], dubbed as **TRBA** which consists of a thin-plate-spline [9] **T**ransformation module, a **R**esNet-based feature extraction network as used in [4], two **B**iLSTM layers with 256 hidden units per layer to converts visual features to contextual sequence of features, and lastly an **A**ttention based LSTM sequential decoder with the hidden state dimension of 256 to convert the sequential features to machine-readable text. Additionally, in the ResNet backbone we introduce dropout layers for Monte-Carlo sampling as depicted in Table 1[1].

RAM	CPU	VRAM	GPU
251 GB	Intel Core i9-10940X	11×4 GB	Nvidia RTX 2080Ti

Table 2. Configuration of the system used to train the models

[1]Implementation based on: https://github.com/clovaai/deep-text-recognition-benchmark

C. Training and Evaluation Details

To train the models we use the AdaDelta [26] optimizer with a learning rate of 1 and a decay rate of $\rho = 0.95$. Furthermore, in total we perform 4 pseudo-label based fully-supervised re-training of the model in a bootstrapped fashion after the initial fully-supervised training with the partially labeled dataset. For each of the fully-supervised training, we train the model for 100K iterations with a batch size of 192. Furthermore, for stable training we use gradient clipping of magnitude 5. Moreover, we use He's method to initialize all parameters. All the models were trained on a single GPU on a server with the configuration described in 2. Algorithm 1 describes the the pseudo-label assignment and selection of the unlabelled data samples, that return $\mathcal{D}_{train}$ updated with the pseudo-labeled samples.

Also, MC-Dropout [6] is notorious for being computationally inefficient since it requires passing every input to each of the sampled model to compute the uncertainty. However, in our implementation we utilize an efficient batch implementation that can easily replace the vanilla Dropout layers in PyTorch[2]. The efficient dropout layer keeps a set of dropout masks fixed while scoring the pool set and exploit batch parallelization for scalibility [12], thus, alleviating the need to pass the input multiple times and the explicit sampling of the models in the ensemble, thus making the system both memory and computationally efficient.

To train the handwriting recognition model we utilize the training splits of the IAM [15] and the CVL [13] and for the scene-text recognition model, contrary to the previous works [1, 2] that use synthetic datasets [7, 8], we use a combination of multiple real scene-text datasets, following the work in [3], for training, that include: IC15 [10], IC13 [11], IIIT [16], SVT [25], SVTP [18], CUTE [19], COCO-Text [24], RCTW [20], Uber-Text [28], ArT [5], MLT [17], and ReCTS [27] consolidating a total of 276k processed images in the training set[3].

The models are evaluated on the IAM [15] and CVL [13] test sets for handwriting recognition. For scene-text recognition we benchmark on six scene-text datasets: IC13 [11], IC15 [10], IIIT [16], SVT [25], SVTP [18], CUTE [19]. For comparison, we also determine the total accuracy, which is the accuracy of the six benchmark datasets combined. Specifically, for scene-text evaluation, the accuracy is calculated only on alphabet and digits, after removing non-alphanumeric characters and normalizing alphabet to lower case. Furthermore, we execute three trials with different seed values for the experiments and report the averaged accuracies.

[2]https://blackhc.github.io/batchbald_redux/consistent_mc_dropout.html

[3]Preprocessed scene-text dataset with the training, validation, and test splits are made available by the authors of [3] at: https://github.com/ku21fan/STR-Fewer-Labels

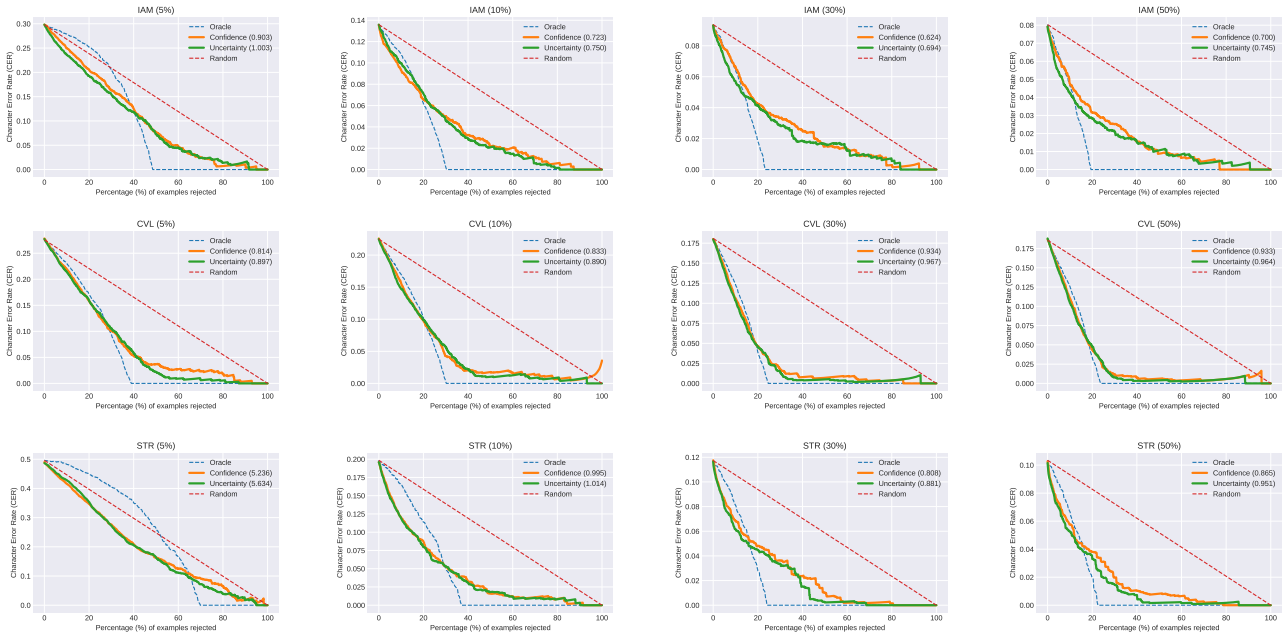

Figure 1. Prediction Rejection Curves w.r.t Character Error Rate (CER). Values in parenthesis in the legend field represent the Prediction Rejection Ratio (PRR) of the corresponding curve.

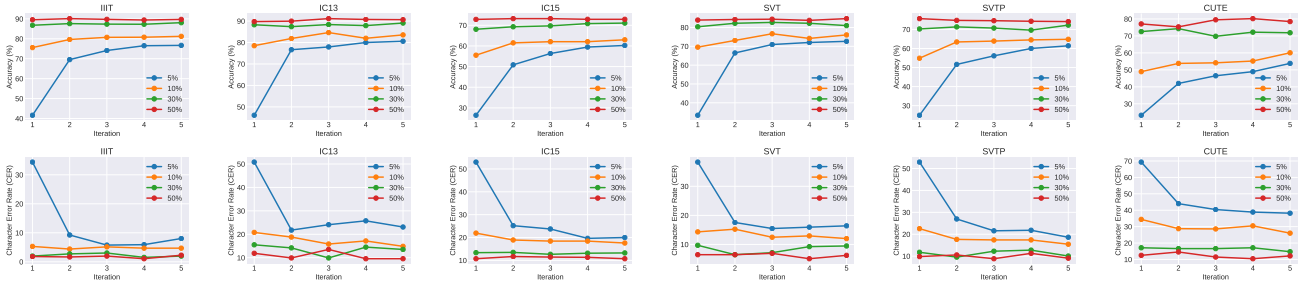

Figure 2. Iteration-wise metric trends of the pseudo-labeling based semi-supervised learning methodology with randomly initialized weights of individual benchmark scene-text datasets.

D. Additional Results

In Figure 1, we visualize the prediction rejection curves w.r.t to the character error rate (CER) of the baseline text recognition model trained on different portions of labeled data on the handwriting and the scene-text datasets.

In Figure 2, we show our vanilla PL-SSL method's performance on word prediction accuracy and CER at the end of each supervised training iteration, starting with different portions of labeled training dataset, for each individual scene-text benchmarks.

Moreover, We conduct experiments with all the text images in the labeled set (276K instances) and the text instances from the TextVQA dataset [21] (463K instances) as the unlabeled set, in Table 3. We found our methods to give on par and in some cases better performance to SeqCLR [1].

References

[1] Aviad Aberdam, Ron Litman, Shahar Tsiper, Oron Anschel, Ron Slossberg, Shai Mazor, R Manmatha, and Pietro Perona. Sequence-to-sequence contrastive learning for text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[2] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[3] Jeonghun Baek, Yusuke Matsui, and Kiyoharu Aizawa. What if we only use real datasets for scene text recognition? toward scene text recognition with fewer labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[4] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang

Method	IIIT5K	SVT	IC13	IC15	SVTP	CUTE80	Total
Supervised Baseline	92.27	85.63	92.22	74.96	75.97	83.68	85.36
SeqCLR (All-to-instance)	92.23	86.86	91.43	76.97	77.99	82.99	86.02
SeqCLR (Frame-to-instance)	91.13	87.79	92.02	77.85	78.30	**86.11**	86.13
SeqCLR (Window-to-instance)	91.23	87.64	**93.01**	**77.90**	80.16	85.76	86.44
Ours	92.73	**88.56**	92.22	76.87	77.98	84.37	86.50
Ours w/ SeqCLR (All-to-instance)	92.33	87.17	91.82	77.85	78.92	85.42	86.55
Ours w/ SeqCLR (Frame-to-instance)	92.83	86.71	92.61	77.36	79.23	**86.11**	86.73
Ours w/ SeqCLR (Window-to-instance)	**93.13**	86.71	91.72	76.83	**81.40**	85.90	**86.76**

Table 3. Word level accuracy (Acc %) using all the labeled data and additional unlabelled data.

Algorithm 1: Pseudo-label assignment and selection of unlabelled data samples at the end of I-th training iteration for the subsequent iteration of supervised training.

Data: $\mathcal{D}_{train}$, $\mathcal{D}_u$, $\boldsymbol{\theta}_I$, τ
Result: $\mathcal{D}_{train}$
N_u = Number of samples in $\mathcal{D}_u$.
$\mathcal{B}_i$ = Hypotheses set for i-th sample.
$\tilde{Y}_i^u$ = Pseudo label for i-th sample.
$i = 1$;
while $i \leq N_u$ **do**
  $\mathcal{B}_i \leftarrow$ beam-search-inference $(f_{\boldsymbol{\theta}_I}(X_i^u)), X_i^u \in \mathcal{D}_u$; /* Deterministic Inference */
  $\tilde{Y}_i^u \leftarrow \arg\max_{Y_i^{(b)}} \{P(Y_i^{(b)}|X_i^u; \boldsymbol{\theta}_I)\}_{b=1}^{B}, Y_i^{(b)} \in \mathcal{B}_i$; /* Pseudo-Label assignment */
  Compute $\mathcal{U}(X_i^u, \mathcal{B}_i)$ using (6) from the main script; /* Stochastic Inference */
  if $\mathcal{U}(X_i^u, \mathcal{B}_i) \leq \tau$ **then**
    $\mathcal{D}_{train} \leftarrow \mathcal{D}_{train} \cup \{X_i^u, \tilde{Y}_i^u\}$;
  end
  i+=1;
end

Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017. 2

[5] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019. 2

[6] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, 2016. 2

[7] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[8] M Jaderberg, K Simonyan, A Vedaldi, and A Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 2

[9] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 2

[10] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, Faisal Shafait, Seiichi Uchida, and Ernest Valveny. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 2015. 1, 2

[11] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazàn Almazàn, and Lluís Pere de las Heras. Icdar 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition (ICDAR)*, 2013. 1, 2

[12] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in Neural Information Processing Systems*, 2019. 2

[13] Florian Kleber, Stefan Fiel, Markus Diem, and Robert Sablatnig. Cvl-database: An off-line database for writer retrieval, writer identification and word spotting. In *2013 12th International Conference on Document Analysis and Recognition (ICDAR)*, 2013. 1, 2

[14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 1

[15] Urs-Viktor Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5:39–46, 2002. 1, 2

[16] Anand Mishra, Alahari Karteek, and C. V. Jawahar. Scene text recognition using higher order language priors. In *British Machine Vision Conference (BMVC)*, 2012. 1, 2

[17] 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)*, 2019. 2

[18] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2013. 1, 2

[19] Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 41(18):8027–8048, 2014. 1, 2

[20] Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017. 2

[21] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[22] Yipeng Sun, Jiaming Liu, Wei Liu, Junyu Han, Errui Ding, and Jingtuo Liu. Chinese street view text: Large-scale chinese text reading with partially supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[23] Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 2019. 2

[24] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016. 1, 2

[25] Kai Wang, Boris Babenko, and Serge J. Belongie. End-to-end scene text recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2011. 1, 2

[26] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012. 2

[27] 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)*, 2019. 2

[28] Ying Zhang, Lionel Gueguen, Ilya Zharkov, Peter Zhang, Keith Seifert, and Ben Kadlec. Uber-text: A large-scale dataset for optical character recognition from street-level imagery. In *SUNw: Scene Understanding Workshop - CVPR 2017*, 2017. 2