Stratified Domain Adaptation: A Progressive Self-Training Approach for Scene Text Recognition
- Supplementary Material -

Kha Nhat Le, Hoang-Tuan Nguyen, Hung Tien Tran, Thanh Duc Ngo*
University of Information Technology, VNU-HCM, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{20520208, 20520344, 19521587}@gm.uit.edu.vn, thanhnd@uit.edu.vn

1. Dataset Descriptions

Our approach leverages *labeled synthetic data* and *unlabeled real data*, as shown in Tab. 1. We **discard the labels** of real datasets to align with the experiments. The "Train." data we report is slightly different from [1, 2] because we use raw images (with discarded labels).

We present some data from the source domain (synthetic) in Fig. 1. Compared to the target domain in Fig. 7, a significant domain gap appears between the two domains, affecting the performance of the STR models.

2. Domain Discriminator (DD) details

2.1. Training detail (stage 1)

Domain Discriminator (DD) employs a binary classifier $f(\boldsymbol{x};\phi)$ with a feature extractor from the baseline model combined with a fully connected layer at the last layer. DD is trained with raw images from S (assigned as class 0) and T (assigned as class 1).

We use focal loss [10] to optimize the learnable parameter to improve DD's accuracy in classifying challenging cases and addressing data imbalance issues (*e.g.* class 0 with 16 million samples and class 1 with 2 million data samples):

$$L(\phi) = -\frac{1}{|S|}\sum_{\boldsymbol{x}^S\in S}(\sigma(f(\boldsymbol{x}^S;\phi)))^{\gamma}\log(1-\sigma(f(\boldsymbol{x}^S;\phi))) - \frac{1}{|T|}\sum_{\boldsymbol{x}^T\in T}(1-\sigma(f(\boldsymbol{x}^T;\phi)))^{\gamma}\log(\sigma(f(\boldsymbol{x}^T,\phi))) \tag{1}$$

where σ is the sigmoid function. Then, we assign $d_i = \sigma(f(\boldsymbol{x}_i;\phi)), d_i \in (0,1)$ to a data point $\boldsymbol{x}_i^T$. The focusing hyper-parameter γ smoothly adjusts the rate at which easy examples are down-weighted.

Figure 1. Examples of synthetic data. The samples are extracted from the MJ and ST datasets.

*Corresponding author

Table 1. Summary of dataset usage. Numbers indicate how many samples were used from each dataset. "t" refers to splits that were repurposed as training data. "*" note that we use the Union14M-Benchmark, which comprises: Artistic, Contextless, Curve, and General.

Dataset	Conf.	Year	# of word boxes		
			Train.	Val.	Eval.
Synthetic datasets					
MJ [5]	NIPSW	2014	7,224,586	802,731[t]	891,924[t]
ST [4]	CVPR	2016	6,975,301	-	-
Real datasets					
IIIT5k [12]	BMVC	2012	2,000	-	3,000
SVT [21]	ICCV	2011	257	-	647
IC13 [8]	ICDAR	2013	848	-	1,015
IC15 [7]	ICDAR	2015	4,468	-	2,077
SVTP [14]	ICCV	2013	-	-	645
CUTE [15]	ESWA	2014	-	-	288
COCO [19]	arXiv	2016	59,820	13,415	9,825
Uber [23]	CVPRW	2017	91,978	36,136	80,418
ArT [3]	ICDAR	2019	32,349	-	35,149
ReCTS [22]	ICDAR	2019	25,328	-	2,592
LSVT [18]	ICDAR	2019	43,244	-	-
MLT19 [13]	ICDAR	2019	56,937	-	-
RCTW17 [16]	ICDAR	2017	10,509	-	-
TextOCR [17]	ECCV	2020	714,770	107,722	-
OpenVINO [9]	ACML	2021	1,914,425	158,819	-
Union14M-Benchmark* [6]	ICCV	2023	-	-	403,379

2.2. Ablation Study on DD (stage 2)

We experimented with the method $\mathrm{StrDA}_{\mathrm{DD}}$ using various settings for the hyper-parameter n. As shown in Fig. 2, Fig. 4, and Fig. 5, in most cases, $\mathrm{StrDA}_{\mathrm{HDGE}}$ demonstrates superior performance compared to $\mathrm{StrDA}_{\mathrm{DD}}$. Moreover, as hyper-parameter n is too high, the effectiveness of StrDA decreases. Therefore, a reasonable choice of n is crucial.

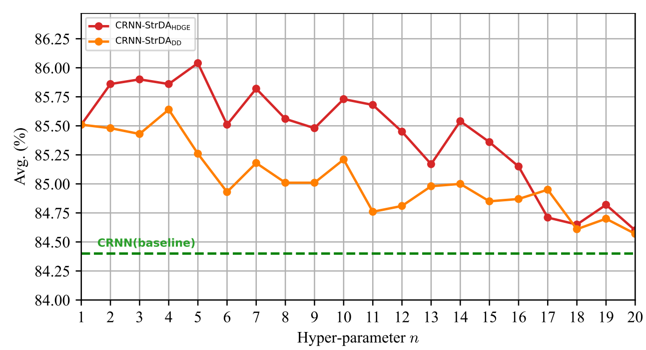

Figure 2. Ablation study on the hyper-parameter n for CRNN-$\mathrm{StrDA}_{\mathrm{HDGE}}$ and CRNN-$\mathrm{StrDA}_{\mathrm{DD}}$.

3. Qualitative Results

In Fig. 6, we visualize the performance of the STR models during the progressive self-training process. $\mathrm{StrDA}_{\mathrm{HDGE}}$ shows improved performance, and the stability of the STR models is reinforced throughout each round of progressive self-training.

In Fig. 3, we observe the predictions of the TRBA-$\mathrm{StrDA}_{\mathrm{HDGE}}$ model in some cases from benchmark datasets. After progressive self-training, the TRBA model gradually improves its accuracy compared to the previous round.

To visually observe how StrDA operates, we sampled some cases from each subset after partitioning. As illustrated in Fig. 7, the difficulty of challenging cases increases gradually through each round. Therefore, when applying progressive self-training to the TRBA model, the recognizer can adapt progressively across each subset from the source to the target domain. $\mathrm{StrDA}_{\mathrm{HDGE}}$ also demonstrates superior performance in generating high-quality pseudo-labels compared to vanilla self-training ST.

4. Limitations and Future Work

A limitation of the proposed method is its dependency on the available target domain data, which is inevitably insufficient to fully cover the target domain. Consequently, if a large portion of the data shares similar patterns, the out-of-distribution (OOD) evaluation will primarily reflect the OOD performance of that specific group. Recently, there has been growing interest in OOD evaluation based on vision foundation models (VFMs) [11, 20]. Utilizing VFMs could provide more generalized output scores.

Moreover, grouping subsets with equal sizes does not accurately reflect the distribution of the domain gap, highlighting the need for a more comprehensive global solution.

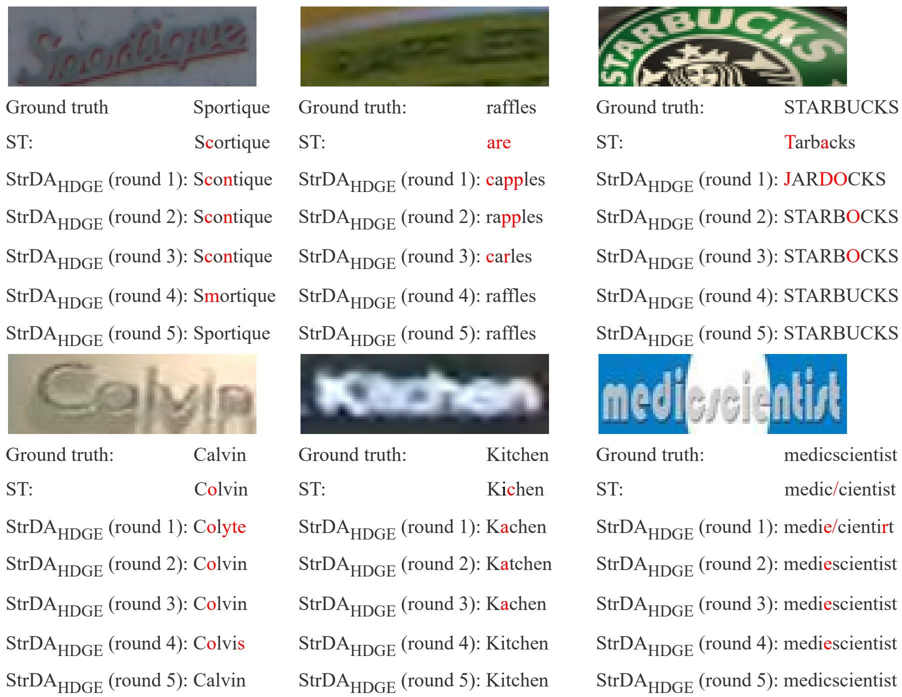

Figure 3. Predictions of TRBA-StrDA$_{\text{HDGE}}$ model on some cases from the benchmark dataset after each round of self-training. It can be seen that the model gradually improves its accuracy compared to the previous round. Misclassified characters are highlighted in red.

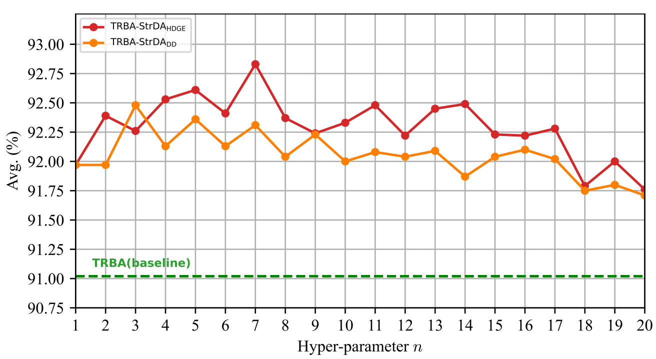

Figure 4. Ablation study on the hyper-parameter n for TRBA-StrDA$_{\text{HDGE}}$ and TRBA-StrDA$_{\text{DD}}$.

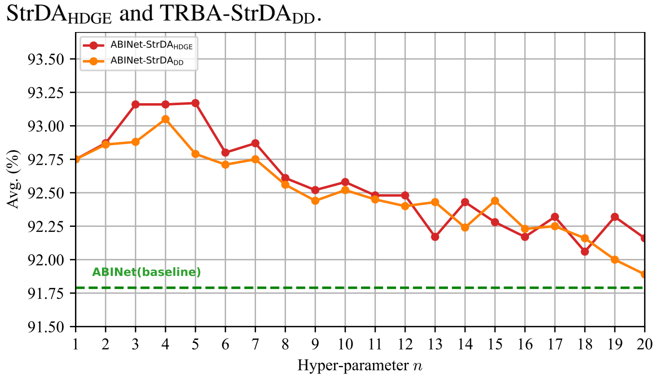

Figure 5. Ablation study on the hyper-parameter n for ABINet-StrDA$_{\text{HDGE}}$ and ABINet-StrDA$_{\text{DD}}$.

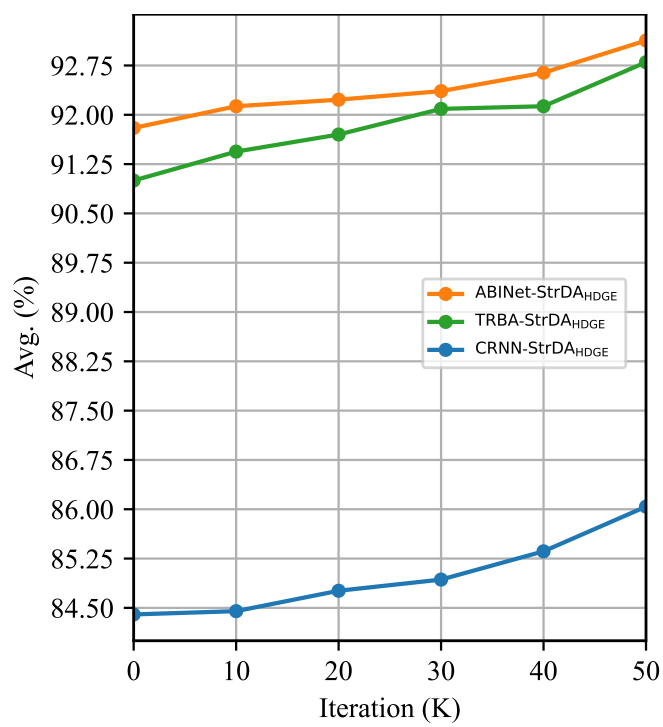

Figure 6. The stability of the STR models throughout the progressive self-training process. It can be observed that the accuracy of the TRBA model steadily increases across rounds.

Figure 7. The Stratified Domain Adaptation ($\mathrm{StrDA}_{\mathrm{HDGE}}$) approach partitions the data from the target domain into five distinct subsets, with the disparity across domains gradually increasing, as shown in the image. The difficulty of challenging cases (curved or perspective texts, occluded texts, texts in low-resolution images, and texts written in difficult fonts) increases progressively across these subsets. The subsets are then subjected to self-training in sequential rounds. We observe the pseudo-labels generated by the TRBA model for each subset at the beginning of the self-training process. In the case of vanilla self-training (ST), all cases are predicted simultaneously by the source-trained (baseline) model. In $\mathrm{StrDA}_{\mathrm{HDGE}}$, the model predicts pseudo-labels for the target domain in round m using the TRBA model after self-training in round $m-1$. The pseudo-labels generated by ST are prone to noise (red characters) as the extent of the domain gap escalates. On the other hand, $\mathrm{StrDA}_{\mathrm{HDGE}}$ produces pseudo-labels with higher quality. This contributes to making the progressive self-training process much more effective. The STR model used for the example is TRBA.

References

[1] 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*, pages 3113–3122, 2021. 1

[2] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *European Conference on Computer Vision*, pages 178–196. Springer, 2022. 1

[3] 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)*, pages 1571–1576. IEEE, 2019. 2

[4] 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. 2

[5] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014. 2

[6] Qing Jiang, Jiapeng Wang, Dezhi Peng, Chongyu Liu, and Lianwen Jin. Revisiting scene text recognition: A data perspective. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20543–20554, 2023. 2

[7] 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

[8] 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. 2

[9] Ilya Krylov, Sergei Nosov, and Vladislav Sovrasov. Open images v5 text annotation and yet another mask text spotter. In *Asian Conference on Machine Learning*, pages 379–389. PMLR, 2021. 2

[10] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1

[11] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022. 2

[12] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC-British machine vision conference*. BMVA, 2012. 2

[13] 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)*, pages 1582–1587. IEEE, 2019. 2

[14] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 569–576, 2013. 2

[15] 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. 2

[16] 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)*, volume 1, pages 1429–1434. IEEE, 2017. 2

[17] 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

[18] 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)*, pages 1557–1562. IEEE, 2019. 2

[19] 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. 2

[20] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023. 2

[21] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *2011 International conference on computer vision*, pages 1457–1464. IEEE, 2011. 2

[22] 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

[23] 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*, volume 2017, page 5, 2017. 2