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

Kha Nhat Le, Hoang-Tuan Nguyen, Hung Tien Tran, Thanh Duc Ngo

{20520208, 20520344, 19521587}@gm.uit.edu.vn, thanhnd@uit.edu.vn

VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Code available !!!

Introduction

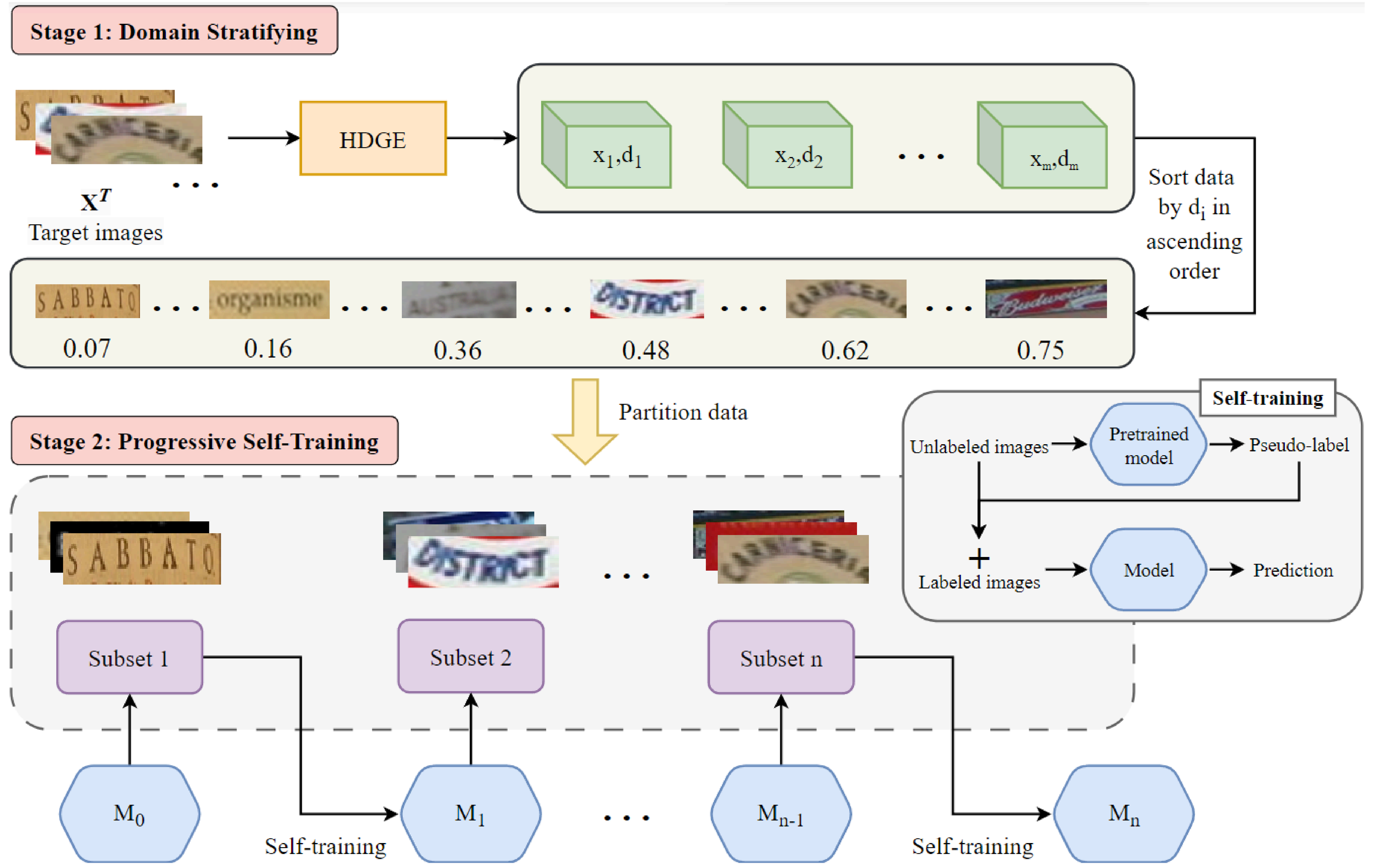
- **Stratified Domain Adaptation (StrDA)**: A two-stage framework utilizing domain stratification by analyzing the gradual escalation of domain gaps, and progressive self-training, **adaptable beyond Scene Text Recognition (STR) tasks**.
- Demonstrates significant improvements across various existing STR models through extensive experiments.
- Offers a cost-effective approach to text recognition, particularly in cases where labeled real data is limited.

Motivation:

- **Data Issue in STR:** The scarcity of labeled real-world data, which is costly and time-intensive to collect.
- **Domain Gap:** STR learning models trained on synthetic datasets achieve notable success; however, their performance often deteriorates when applied to real-world data
- **Limitations of Previous Methods:** Direct unsupervised domain adaptation (UDA) often fail in bridging significant domain gaps.

Method

Overall framework of our proposed StrDA for STR



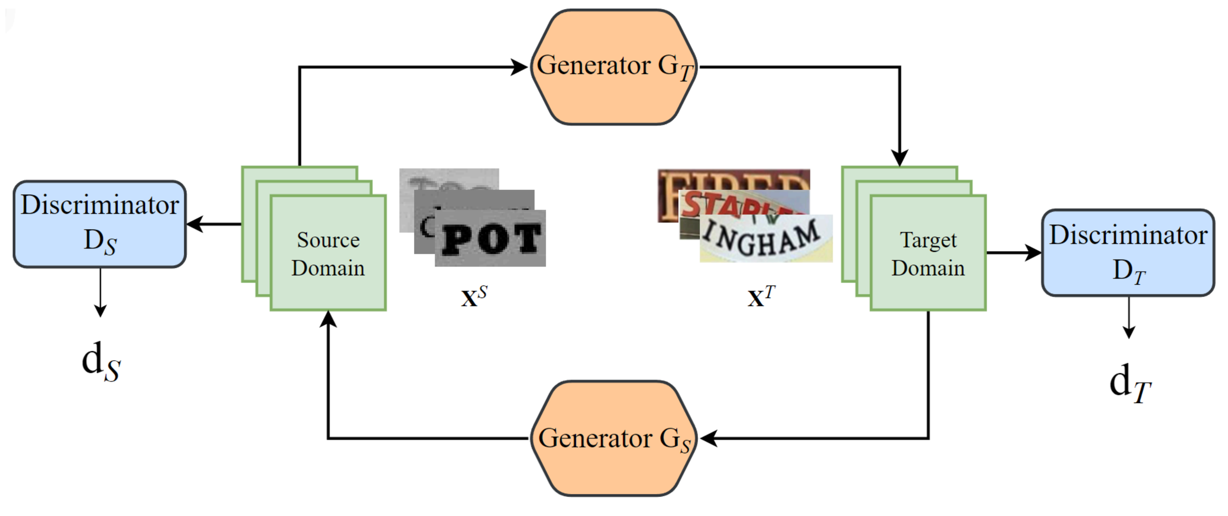
Harmonic Domain Gap Estimator (HDGE):

are trained under the adversarial manner

* Source domain $S = (x_i^S, y_i^S)_{i=1}^{|S|}$ and target domain $T = (x_i^T)_{i=1}^{|T|}$

Partition into a series of equally-sized groups $T_m = (x_i^{T_m})_{i=1}^{|T_m|}$:

$$\rho(S, T_m) \le \rho(S, T_{m+1}) \quad \forall m \in (1, n) \qquad (1)$$



Generators $G_T: S \rightarrow T$, $G_S: T \rightarrow S$; **Discriminators** D_S, D_T

$$d_i = \frac{(1+\beta^2).\, d_S(x_i^T).\, d_T(x_i^T)}{\beta^2.\, d_S(x_i^T) + d_T(x_i^T)} \qquad (2)$$

$d_S(x_i^T)$ and $d_T(x_i^T)$ represent out-of-distribution (OOD) level of x_i^T with respect to source domain and target domain, respectively

Ablation Study

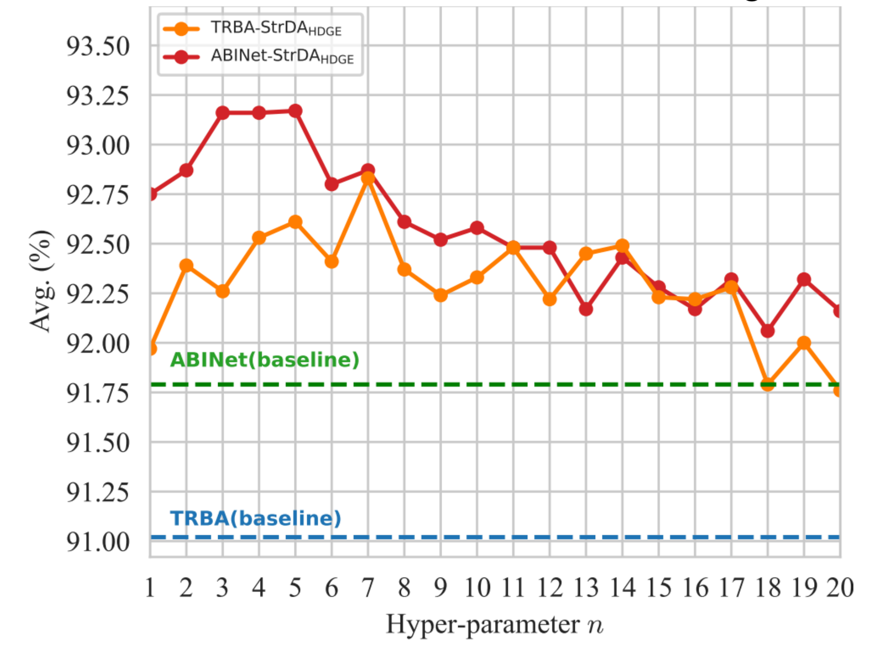


Figure 3. Ablation study on Hyper-parameter n in Eq. (1)

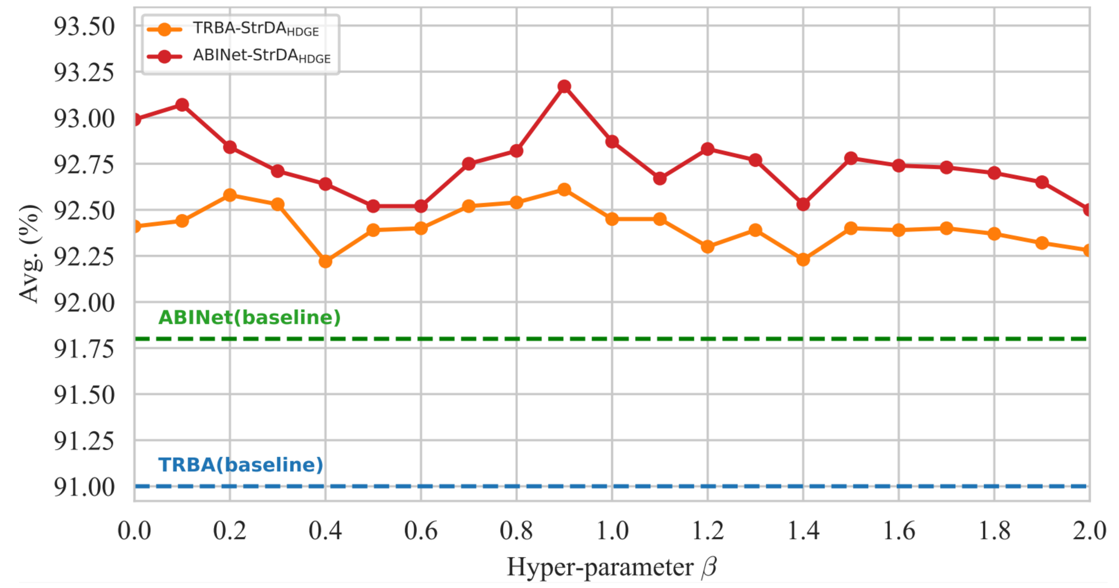


Figure 4. Ablation study on Hyper-parameter β in Eq. (2)

Quantitative Results

Table 1. Word accuracy on six scene-text benchmarks and five additional datasets. "ST" refers to traditional unsupervised domain adaptation (vanilla self-training)

Type	Method	Common Benchmarks							Additional Datasets				
		IIIT 3,000	SVT 647	IC13 857	IC15 1,811	SVTP 645	CUTE 288	Avg.	COCO 9,825	Uber 80,418	ArT 35,149	ReCTS 2,592	Union14M 403,379
CTC	CRNN (baseline)	92.5	86.4	92.0	71.3	75.2	83.3	84.4	49.5	34.6	59.5	77.1	43.3
	+ ST	93.7	87.6	92.2	72.9	75.5	84.7	85.5	51.4	35.9	60.7	79.8	46.2
	Δ	+1.2	+1.2	+0.2	+1.6	+0.3	+1.4	+1.1	+1.9	+1.3	+1.2	+2.7	+2.9
	+ StrDA$_{HDGE}$	93.4	89.0	93.1	74.0	77.1	84.4	86.0	53.0	36.8	60.9	81.0	47.8
	Δ	+0.9	+2.6	+1.1	+2.7	+1.9	+1.1	+1.6	+3.5	+2.2	+1.4	+3.9	+4.5
Attention	TRBA (baseline)	96.2	93.7	95.8	81.9	86.1	91.0	91.0	62.5	39.0	69.0	82.8	56.6
	+ ST	97.1	94.0	96.1	82.5	90.1	92.4	92.0	65.5	40.9	70.9	84.8	60.4
	Δ	+0.9	+0.3	+0.3	+0.6	+4.0	+1.4	+1.0	+3.0	+1.9	+1.9	+2.0	+3.8
	+ StrDA$_{HDGE}$	97.2	95.2	**96.5**	**84.5**	90.7	**94.4**	92.8	68.6	42.7	**72.2**	**85.8**	**64.2**
	Δ	+1.0	+1.5	+0.7	+2.6	+4.6	+3.4	+1.8	+6.1	+3.7	+3.2	+3.0	+7.6
LM	ABINet (baseline)	97.0	95.2	95.6	82.3	89.5	90.3	91.8	63.2	39.5	68.9	82.6	55.7
	+ ST	97.4	96.3	96.4	83.9	**91.0**	92.0	92.8	68.7	42.3	71.2	84.7	61.4
	Δ	+0.4	+1.1	+0.8	+1.6	+1.5	+1.7	+1.0	+5.5	+2.8	+2.3	+2.1	+5.7
	+ StrDA$_{HDGE}$	**97.8**	**96.9**	96.0	84.4	**91.0**	**94.4**	**93.2**	**69.7**	**44.2**	71.6	85.0	62.9
	Δ	+0.8	+1.7	+0.4	+2.1	+1.5	+4.1	+1.4	+6.5	+4.7	+2.7	+2.4	+7.2

Table 2. Comparison with other domain adaptation methods in the STR task. Our method significantly enhances the performance of the STR models, surpassing other existing approaches. Additionally, it can be integrated with other methods to achieve even greater efficacy

	Method	Labeled Dataset	Unlabeled Dataset	Regular Text: IIIT 3000	SVT 647	IC13 857	IC13 1015	Irregular Text: IC15 1811	IC15 2077	SVTP 645	CUTE 288
Published Results	TRBA-FEDS [41]	MJ+ST	Amazon_book_cover	92.2	92.1	96.5	95.3	83.8	80.9	84.0	79.0
	TRBA-Seq-UPS [40]	MJ+ST	276K RU	92.7	88.6	_	92.2	_	76.9	78.0	84.4
	TRBA-cr [71]	MJ+ST	10.6M RU	96.5	96.3	**98.3**	_	**89.3**	_	**93.3**	93.4
	ABINet-st [16]	MJ+ST	Uber-Text	96.8	94.9	97.3	_	87.4	_	90.1	93.4
	ABINet-est [16]	MJ+ST	Uber-Text	97.2	95.5	97.7	_	86.9	_	89.9	94.1
Our Results	TRBA-cr (reproduce)	MJ+ST	2M RU	97.3	95.1	97.2	96.2	88.1	84.0	90.5	93.8
	TRBA-StrDA$_{HDGE}$	MJ+ST	2M RU	97.2	95.2	97.4	96.5	88.4	**84.5**	90.7	**94.4**
	TRBA-StrDA$_{HDGE}$ w/ cr	MJ+ST	2M RU	97.3	96.1	97.6	**96.7**	88.7	**84.5**	90.9	**94.4**
	ABINet-StrDA$_{HDGE}$	MJ+ST	2M RU	**97.8**	**96.9**	97.0	96.0	88.6	84.4	91.0	**94.4**

Qualitative Results

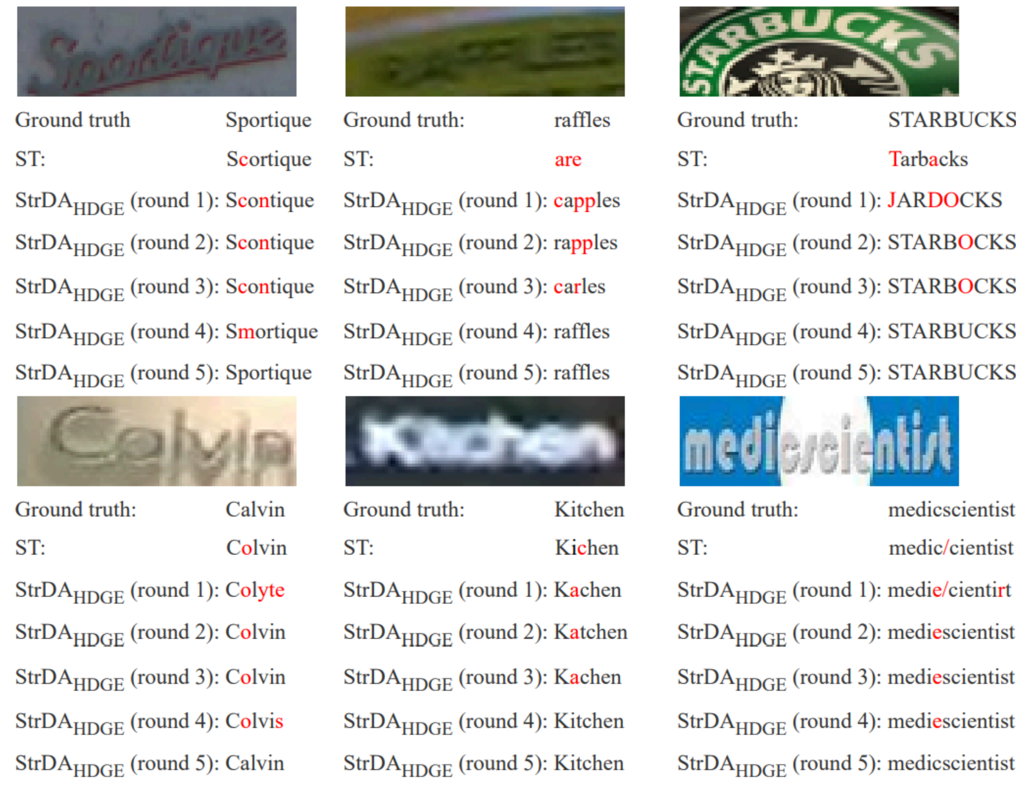


Figure 1. Predictions of TRBA-StrDA model on some cases from the benchmark dataset after each round of self-training

Figure 2. StrDA partitions the data from the target domain into five distinct subsets, with the disparity across domains gradually increasing, as shown in the image

[46] Shi *et al.*, An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, in *IEEE Trans. Pattern Anal. Mach. Intell. 2016*. [3] Baek *et al.*, What is wrong with scene text recognition model comparisions? dataset and model analysis, in *ICCV 2019*. [16] Fang *et al.*, Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition, in *CVPR 2021*. [41] Patel and Matas, Feds-filtered edit disstance surrogate, in *ICDAR 2021*. [40] Patel *et al.*, Seq-ups: sequential uncertainty-aware pseudo-label selection for semi-supervised text recognition, in *WACV 2023*. [71] Zheng *et al.*, Pushing the performance limit of scene text recognizer without human annotation, in *CVPR 2022*.