VIET NAM NATIONAL UNIVERSITY HO CHI MINH CITY

VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT

WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

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

IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2025

Kha Nhat Le
(Presenter)

Hoang-Tuan Nguyen

Hung Tien Tran

Thanh Duc Ngo

University of Information Technology, VNU-HCM, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam

Domain Gap in Scene Text Recognition (STR)

SYNTHETIC DATA
(source domain)

REAL-WORLD DATA *
(target domain)

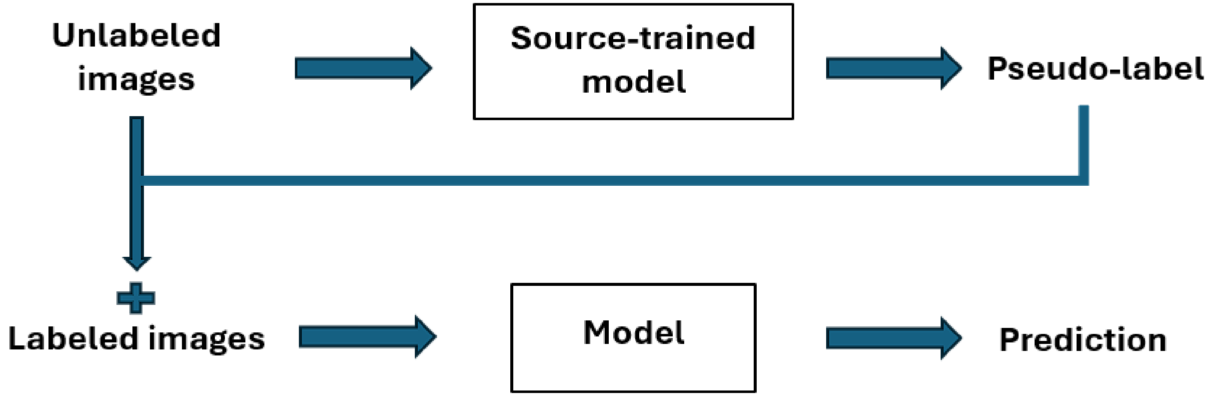
Difficult fonts **Occlusion** **Blur**

Curved text

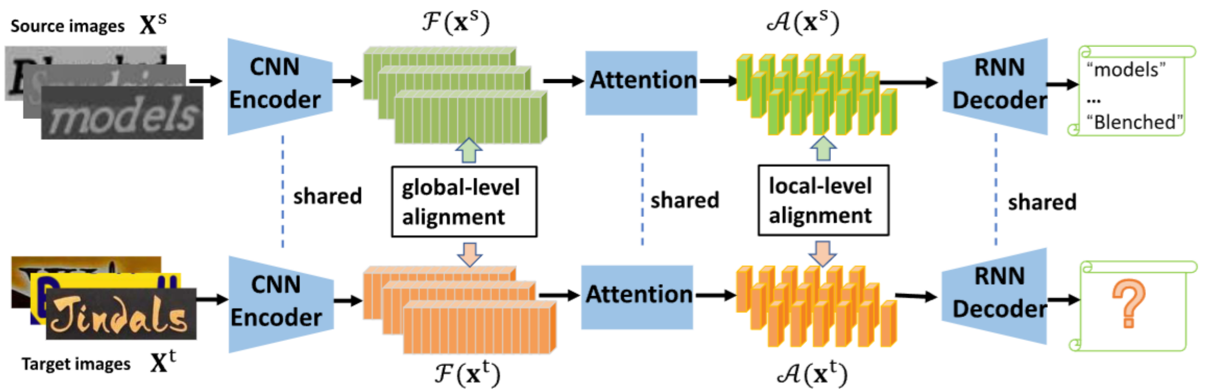
- STR learning models requires huge amounts of labelled data
- Gathering labeled real data is challenging (high cost and time-intensive nature)
- Many models primarily utilize synthetic data for training

(*) Images for real-wrold data are taken from [4] Baek, J., Matsui, Y. and Aizawa, K., 2021. *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* (pp. 3113-3122).

Unsupervised Domain Adaptation (UDA) in STR



Baek *et al.* [4]



Zhang *et al.* [68]

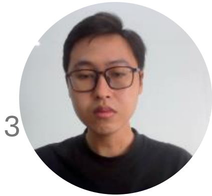
Issue

Large gap between source and target domains

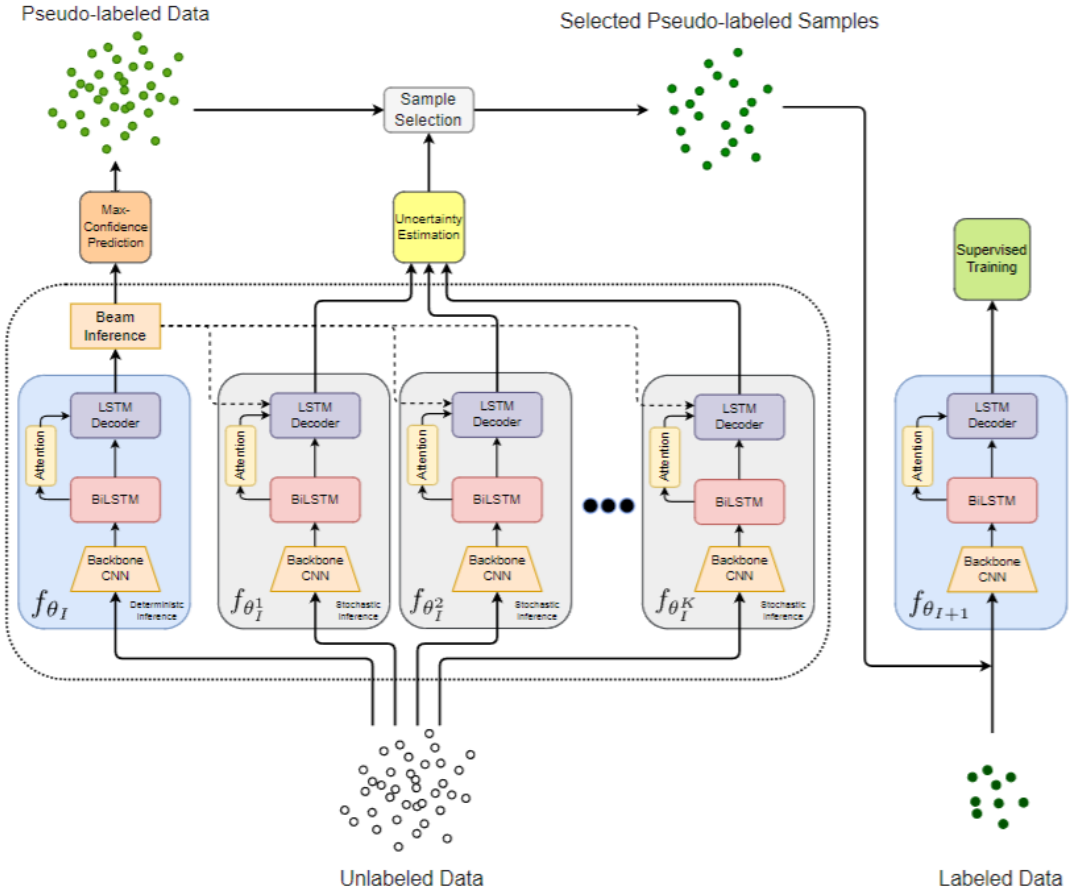
[28]

The efficacy of UDA tends to degrade

[4] Baek, J., Matsui, Y. and Aizawa, K., 2021. 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* (pp. 3113-3122).
[28] Kumar, A., Ma, T. and Liang, P., 2020, November. Understanding self-training for gradual domain adaptation. In *International conference on machine learning* (pp. 5468-5479). PMLR.
[68] Zhang, Y., Nie, S., Liang, S. and Liu, W., 2021. Robust text image recognition via adversarial sequence-to-sequence domain adaptation. *IEEE Transactions on Image Processing, 30*, pp.3922-3933.



Multiple Self-training Steps Strategy



Patel *et al.* [40]

Algorithm 1 Ensemble Self-training

Require: Labeled images $\mathcal{X}$ with labels $\mathcal{Y}$ and unlabeled images $\mathcal{U}$

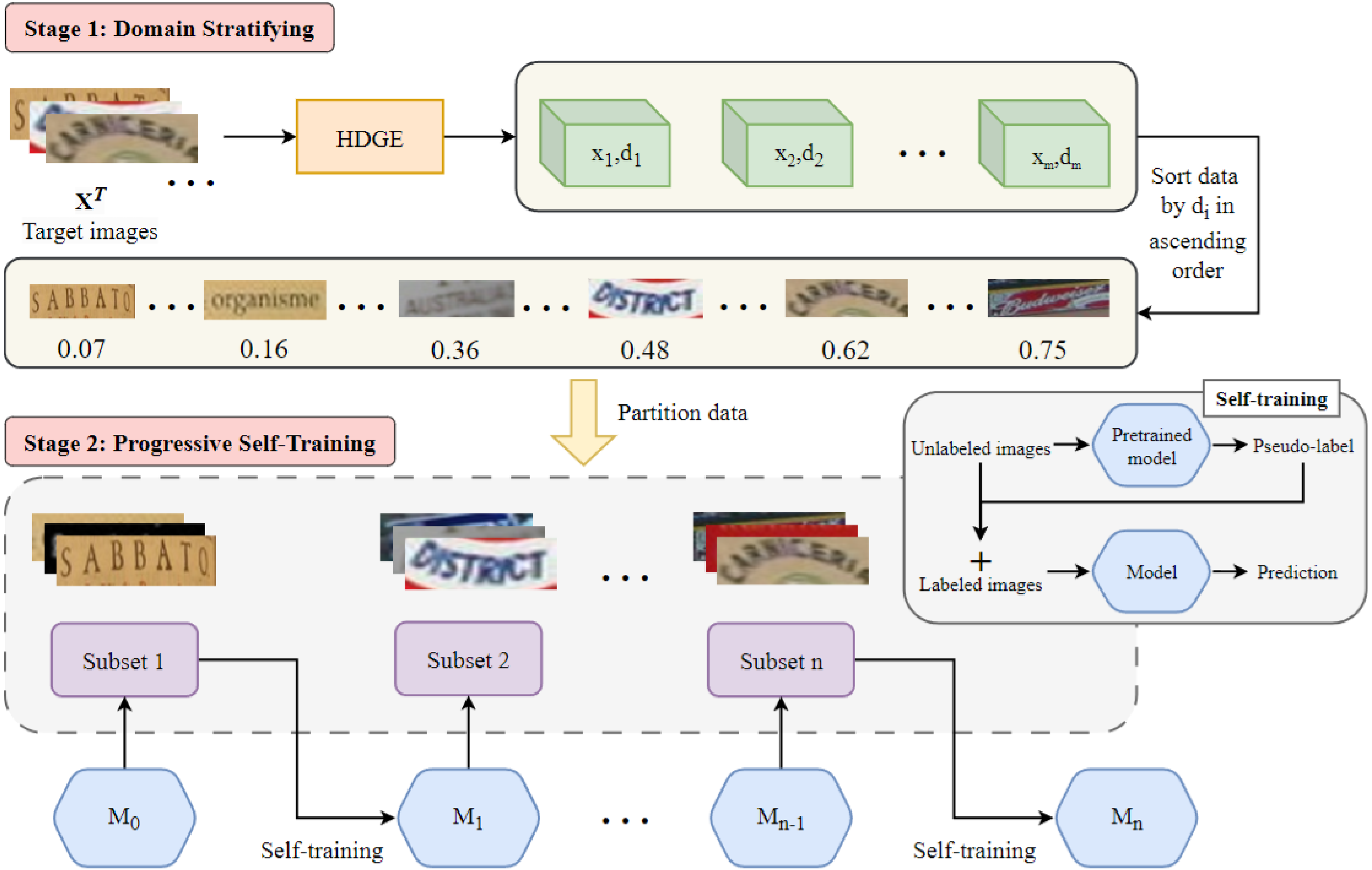
Train parameters θ_0 of ABINet with $(\mathcal{X}, \mathcal{Y})$ using Equation 8.
Use θ_0 to generate soft pseudo labels $\mathcal{V}$ for $\mathcal{U}$
Get $(\mathcal{U}', \mathcal{V}')$ by filtering $(\mathcal{U}, \mathcal{V})$ with $\mathcal{C} < Q$ (Equation 9)
for $i = 1, \ldots, N_{max}$ **do**
  if $i == N_{upl}$ **then**
    Update $\mathcal{V}$ using θ_i
    Get $(\mathcal{U}', \mathcal{V}')$ by filtering $(\mathcal{U}, \mathcal{V})$ with $\mathcal{C} < Q$ (Equation 9)
  end if
  Sample $B_l = (\mathcal{X}_b, \mathcal{Y}_b) \subsetneqq (\mathcal{X}, \mathcal{Y})$, $B_u = (\mathcal{U}'_b, \mathcal{V}'_b) \subsetneqq (\mathcal{U}', \mathcal{V}')$
  Update θ_i with B_l, B_u using Equation 8.
end for

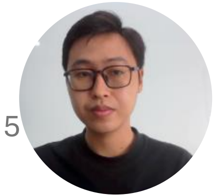
Fang *et al.* [16]

[16] Fang, S., Xie, H., Wang, Y., Mao, Z. and Zhang, Y., 2021. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7098-7107).
[40] Patel, G., Allebach, J.P. and Qiu, Q., 2023. Seq-ups: Sequential uncertainty-aware pseudo-label selection for semi-supervised text recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 6180-6190).

TUCSON, ARIZONA • FEB 28 - MAR 4

Ours: Stratified Domain Adaptation (StrDA)



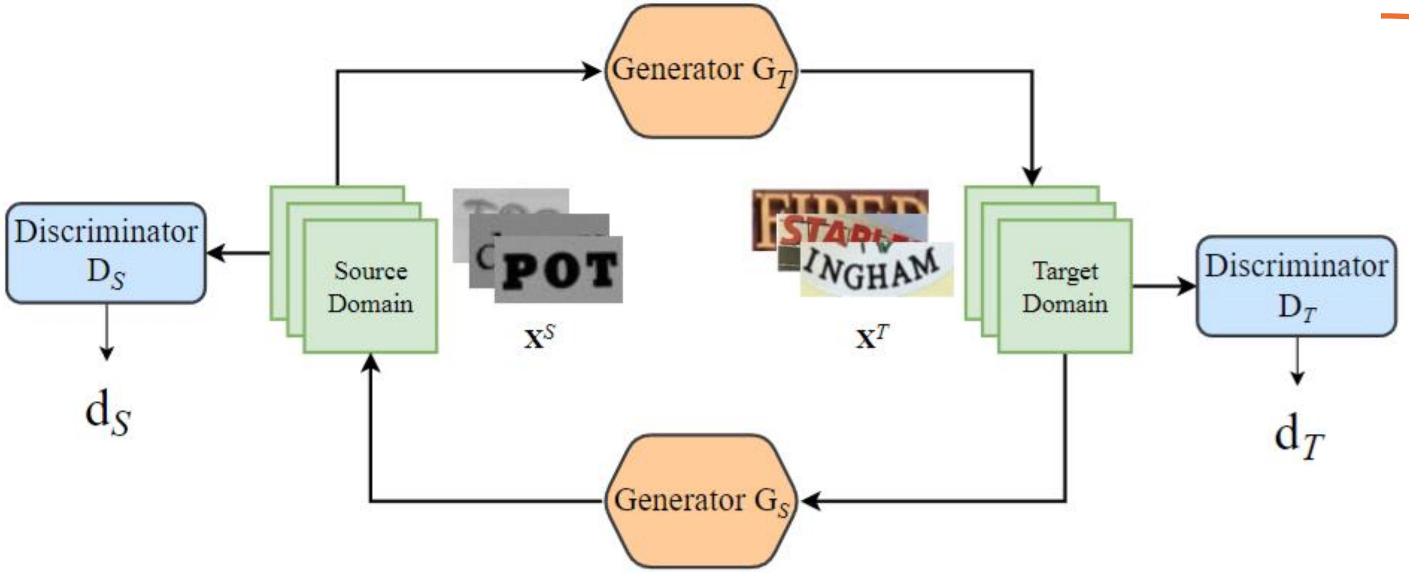


Stage 1: Domain Stratifying

* 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 $T_m = (x_i^{T_m})_{i=1}^{|T_m|}$:

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

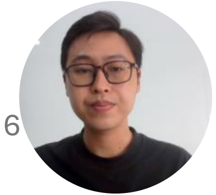


Generators $G_T: S \rightarrow T$, $G_S: T \rightarrow S$; **discriminators** D_S, D_T

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

Harmonic Domain Gap Estimator (HDGE)

$d_S(x_i^T)$ 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



Stage 2: Progressive Self-Training

WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

Algorithm 1 Progressive Self-Training ST

Require: Labeled images $(X, Y) \in S$ and sequence of unlabeled image subsets $T_1, T_2, T_3, \ldots, T_n (T_i \in T)$

```
Train STR model M(·, θ_0) with (X, Y) using CE loss.
for iteration i = 1, 2, ..., n do
    T_i → M(·, θ_{i−1}) → V_i (pseudo-labels) and m_i (average confidence-scores)
    Update θ_i with (X, Y), (T_i, V_i), m_i using Eq. (5)
end for
```

Objective:

$$L(\theta) = \frac{1-m_i}{|S|} \sum_{x^S \in S} L_r\left(x^S; y^S\right) + \frac{m_i}{|T_i|} \sum_{x^{T_i} \in T_i} L_r\left(x^{T^i}; y^{T^i}\right) \quad (5)$$

where m_i is the mean (average) of confidence scores when generating pseudo-labels for the unlabeled image subset T^i. m_i serves as an *adaptive controller*.

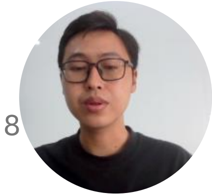
Experiment

- **Datasets:** 16 million labeled synthetic data + 2 million unlabeled real data

- **STR models:** CRNN [46], TRBA [3], and ABINet [16]

- **Evaluation Metrics:** word-accuracy for each dataset

[3] Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S.J. and Lee, H., 2019. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4715-4723).
[16] Fang, S., Xie, H., Wang, Y., Mao, Z. and Zhang, Y., 2021. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7098-7107).
[46] Shi, B., Bai, X. and Yao, C., 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence, 39*(11), pp.2298-2304.



Quantitative Result

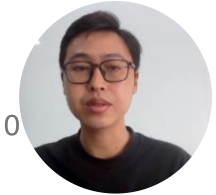
Type	Method	IIIT 3,000	SVT 647	IC13 857	IC15 1,811	SVTP 645	CUTE 288	Avg.	COCO 9,825	Uber 80,418	ArT 35,149	ReCTS 2,592	Union14M 403,379
		Common Benchmarks							Additional Datasets				
CTC	CRNN (baseline)	92.5	86.4	92.0	71.3	75.2	83.3	84.4	49.5	34.6	59.5	77.1	43.3
	+ ST	93.7	87.6	92.2	72.9	75.5	84.7	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
	+ $\text{StrDA}_{\text{HDGE}}$	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
	+ $\text{StrDA}_{\text{HDGE}}$	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
	+ $\text{StrDA}_{\text{HDGE}}$	**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

Comparison with other UDA in STR task

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				Irregular Text			
				IIIT	SVT	IC13		IC15		SVTP	CUTE
				3000	647	857	1015	1811	2077	645	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$_{\text{HDGE}}$	MJ+ST	2M RU	97.2	95.2	97.4	96.5	88.4	**84.5**	90.7	**94.4**
	TRBA-StrDA$_{\text{HDGE}}$ w/ cr	MJ+ST	2M RU	97.3	96.1	97.6	**96.7**	88.7	**84.5**	90.9	**94.4**
	ABINet-StrDA$_{\text{HDGE}}$	MJ+ST	2M RU	**97.8**	**96.9**	97.0	96.0	88.6	84.4	91.0	**94.4**

[16] Fang, S., Xie, H., Wang, Y., Mao, Z. and Zhang, Y., 2021. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7098-7107).
[40] Patel, G., Allebach, J.P. and Qiu, Q., 2023. Seq-ups: Sequential uncertainty-aware pseudo-label selection for semi-supervised text recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 6180-6190).
[41] Patel, Y. and Matas, J., 2021, September. FEDS-filtered edit distance surrogate. In *International Conference on Document Analysis and Recognition* (pp. 171-186). Cham: Springer International Publishing.
[71] Zheng, C., Li, H., Rhee, S.M., Han, S., Han, J.J. and Wang, P., 2022. Pushing the performance limit of scene text recognizer without human annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14116-14125).



Qualitative Result

WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

Conclusion

- We introduce a progressive self-training domain adaptation approach for scene text recognition, which helps improve the model's performance by utilizing unlabeled data with high-quality pseudo-labels.
- This paves the way for recognizing text without incurring human annotation costs, particularly in cases where labeled real data is limited.

Future work

- Utilizing vision foundation model (VFMs) could provide more generalized out-of-distribution (OOD) evaluation
- Find a general method to determine a reasonable amount of data in each sub-domain group (not equal-size for all sub-groups)
- Apply the approach to other problems with higher complexity and larger domain gaps, such as medical image segmentation

THANK YOU

Code available !!!

https://github.com/KhaLee2307/StrDA