SVTRv2: CTC Beats Encoder-Decoder Models in Scene Text Recognition

FUDAN UNIVERSITY
复旦大学可信具身智能研究院
Institute of Trustworthy Embodied AI, Fudan University

Yongkun Du, Zhineng Chen*, Hongtao Xie, Caiyan Jia, Yu-Gang Jiang

zhinchen@fudan.edu.cn; ykdu23@m.fudan.edu.cn

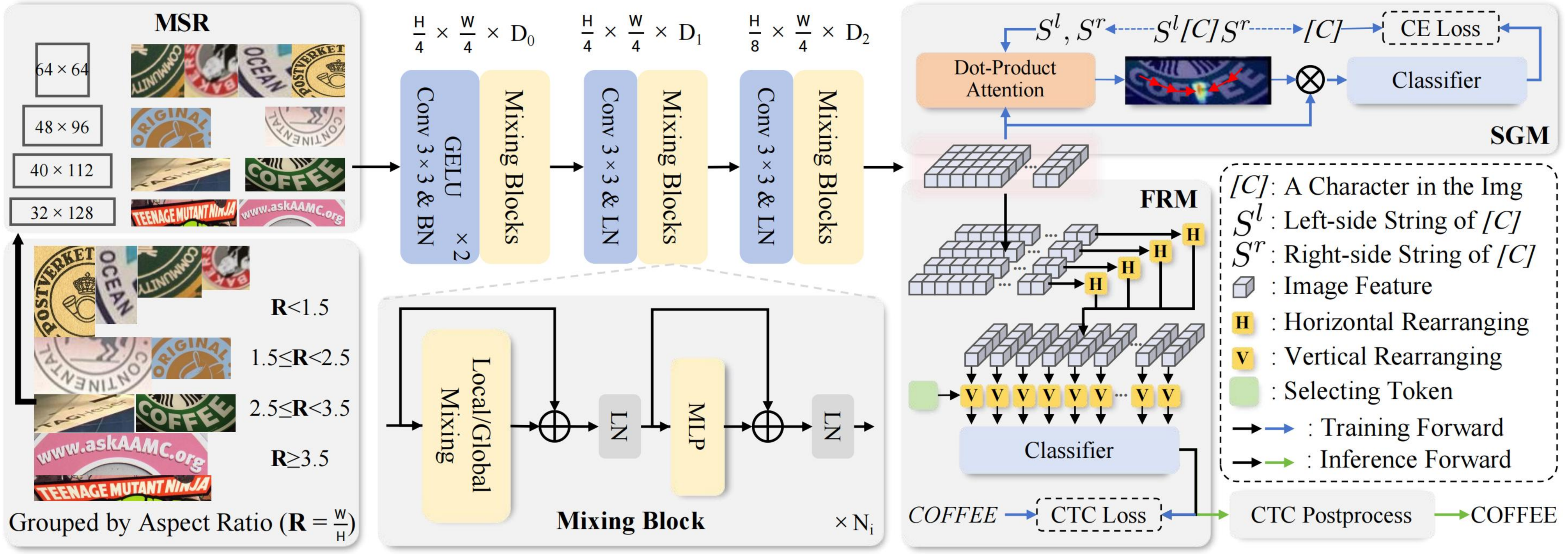
ICCV HONOLULU HAWAII OCT 19-23, 2025

We propose SVTRv2, a CTC model endowed with the ability to handle text irregularities and model linguistic context. First, a Multi-Size Resizing (MSR) strategy is proposed to resize text instances to appropriate predefined sizes, effectively avoiding severe text distortion. Meanwhile, we introduce a Feature Rearrangement Module (FRM) to ensure that visual features accommodate the requirement of CTC, thus alleviating the alignment puzzle. Second, we propose a Semantic Guidance Module (SGM). It integrates linguistic context into the visual features, allowing CTC model to leverage language information for accuracy improvement. Code is available at: https://github.com/Topdu/OpenOCR

Motivation

- As shown in the Figure above. CTC models generally exhibit worse accuracy than encoder-decoder-based methods (EDTRs) due to struggling with text irregularity and linguistic missing.

Method



MSR resizes text images to four predefined sizes based on their aspect ratios (R=W/H), with each size corresponding to a specific R range (R<1.5, 1.5≤R<2.5, 2.5≤R<3.5, R≥3.5) .

It minimizes text distortion caused by fixed-size resizing, maintains the discriminability of the text image and enables the CTC model to handle arbitrary shaped text images

FRM applies horizontal and horizontal rearrangement with MHA to get the final sequence F^v aligned with text reading order.

$$\mathbf{M}^h_i = \sigma\left(\mathbf{F}_i\mathbf{W}^q_i\left(\mathbf{F}_i\mathbf{W}^k_i\right)^t\right)$$
$$\mathbf{F}^{h'}_i = \mathrm{LN}(\mathbf{M}^h_i\mathbf{F}_i\mathbf{W}^v_i + \mathbf{F}_i)$$
$$\mathbf{F}^h_i = \mathrm{LN}(\mathrm{MLP}(\mathbf{F}^{h'}_i) + \mathbf{F}^{h'}_i)$$
$$\mathbf{M}^v_j = \sigma\left(\mathbf{T}^s\left(\mathbf{F}^h_{:,j}\mathbf{W}^k_j\right)^t\right)$$
$$\mathbf{F}^v_j = \mathbf{M}^v_j\mathbf{F}^h_{:,j}\mathbf{W}^v_j$$

SGM guides the visual model to integrate left and right linguistic context into visual features.

$$\mathbf{Q}^l_i = \mathrm{LN}\left(\sigma\left(\mathbf{T}^l\mathbf{W}^q\left(\mathbf{E}^l_i\mathbf{W}^k\right)^t\right)\mathbf{E}^l_i\mathbf{W}^v + \mathbf{T}^l\right)$$
$$\mathbf{A}^l_i = \sigma\left(\mathbf{Q}^l_i\mathbf{W}^q\left(\mathbf{F}\mathbf{W}^k\right)^t\right),\ \mathbf{F}^l_i = \mathbf{A}^l_i\mathbf{F}\mathbf{W}^v$$

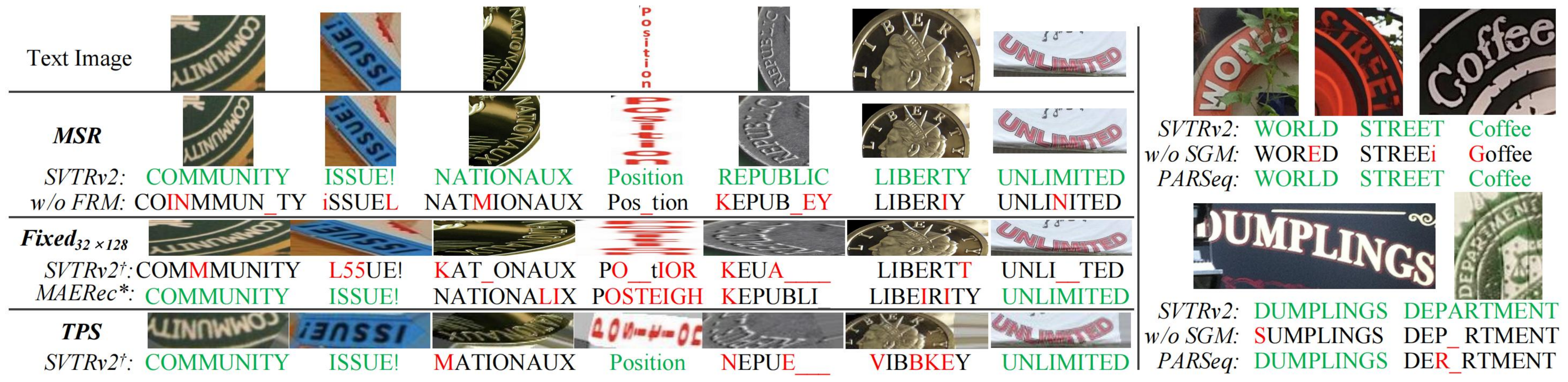
It significantly improving the recognition accuracy in occluded text. It is discarded during inference not increasing the inference time cost.

Ablation Study

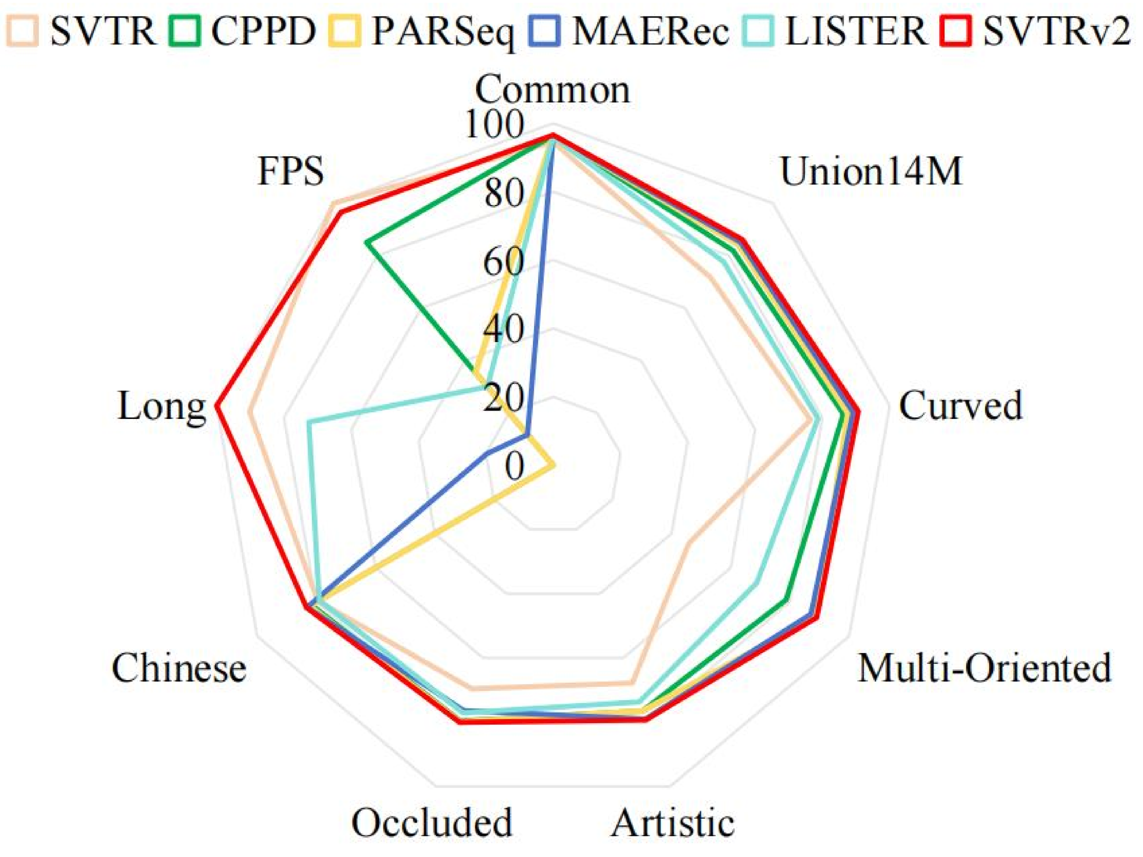
		R_1 2,688	R_2 788	R_3 266	R_4 32	*Curve*	*MO*	*Com*	*U14M*
SVTRv2 (+MSR+FRM)		87.4	88.3	86.1	87.5	88.17	86.19	96.16	83.86
SVTRv2 (w/o both)		70.5	81.5	82.8	84.4	82.89	65.59	95.28	77.78
vs. MSR (+FRM)	$\text{Fixed}_{32\times128}$	72.1	83.1	84.1	85.6	83.18	68.71	95.56	78.87
	$\text{Padding}_{32\times W}$	52.1	71.3	82.3	87.4	71.06	51.57	94.70	71.82
	$\text{Fixed}_{64\times256}$	76.6	81.6	81.9	80.2	85.70	67.49	95.07	79.03
vs. FRM (+MSR)	w/o FRM	85.7	86.3	86.0	85.5	87.35	83.73	95.44	82.22
	+ H rearranging	87.0	87.1	86.3	85.5	88.05	85.76	95.98	82.94
	+ V rearranging	85.0	87.6	88.5	85.5	88.01	84.44	95.66	82.70
	+ TF_1	86.4	86.3	87.5	86.1	87.51	85.50	95.60	82.49

	Method	OST_w	OST_h	Avg	*Com*	*U14M*
Linguistic context modeling	w/o SGM	82.86	66.97	74.92	96.16	83.86
	SGM	**86.26**	**73.80**	**80.03**	**96.57**	**86.14**
	GTC [23]	83.07	68.32	75.70	96.01	84.33
	ABINet [15]	83.07	67.54	75.31	96.25	84.17
	VisionLAN [47]	83.25	68.97	76.11	96.39	84.01
	PARSeq [4]	83.85	69.24	76.55	96.21	84.72
	MAERec [25]	83.21	69.69	76.45	96.47	84.69

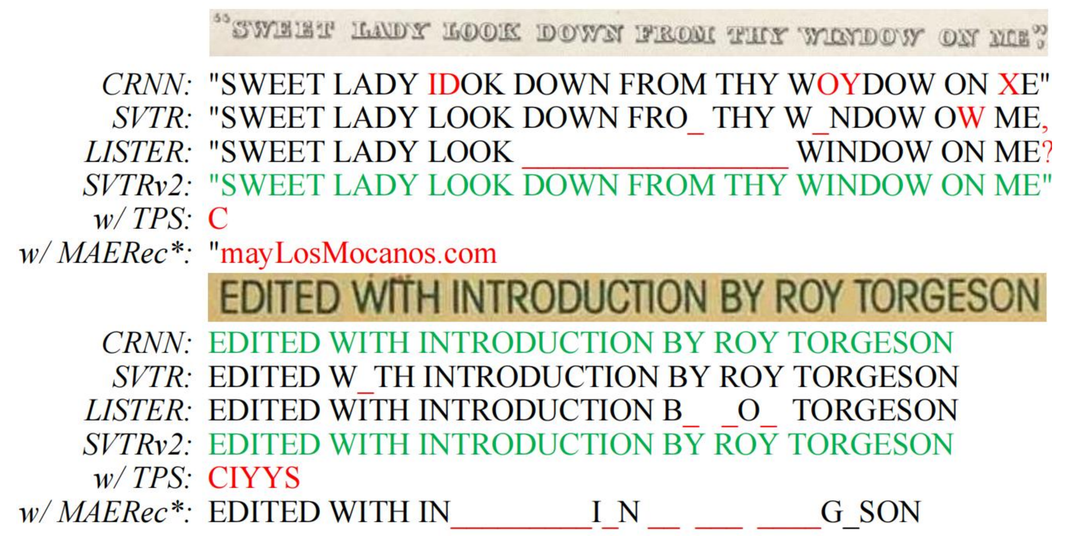
- In scenarios involving Curve and multi-oriented (MO) text, MSR achieves a 15.3% improvement over fixed scaling methods on curved text with R_1 and R_2 .
- FRM enhances recognition accuracy for multi-oriented text by 2.46% and is even capable of correctly recognizing upside-down text.
- In cases of occluded text, SGM demonstrates particularly notable effectiveness, improving accuracy by 5.11%, substantially surpassing the gains achieved by advanced language models such as ABINet and PARSeq.



Comparison with State-of-the-arts



		R_1	R_2	R_3	R_4	*Curve*	*MO*	*Com*	*U14M*	*LTB*
SVTRv2		**90.8**	**89.0**	**90.4**	**91.0**	**90.64**	**89.04**	**96.57**	**86.14**	**50.2**
TPS	SVTR [11]	86.8	82.3	77.3	75.7	82.19	86.12	94.62	78.44	0.0
	SVTRv2	89.5	85.1	78.4	83.8	84.71	88.97	94.62	79.94	0.5
MAE-REC*	SVTR [11]	81.3	87.6	87.6	88.3	87.88	78.74	96.32	83.23	0.0
	SVTRv2	88.0	88.9	89.4	88.3	89.96	87.56	96.42	85.67	0.2



Method	*Scene*	*Web*	*Doc*	*HW*	Avg	$Scene_{L>25}$	*Size*
ASTER [40]	61.3	51.7	96.2	37.0	61.55	-	27.2
MORAN [32]	54.6	31.5	86.1	16.2	47.10	-	28.5
SAR [29]	59.7	58.0	95.7	36.5	62.48	-	27.8
SEED [36]	44.7	28.1	91.4	21.0	46.30	-	36.1
MASTER [31]	62.8	52.1	84.4	26.9	56.55	-	62.8
ABINet [15]	66.6	63.2	98.2	53.1	70.28	-	53.1
TransOCR [5]	71.3	64.8	97.1	53.0	71.55	-	83.9
CCR-CLIP [56]	71.3	69.2	98.3	60.3	74.78	-	62.0
DCTC [60]	73.9	68.5	99.4	51.0	73.20	-	40.8
CAM [54]	76.0	69.3	98.1	59.2	76.80	-	135
PARSeq* [4]	84.2	82.8	**99.5**	63.0	82.37	0.0	28.9
MAERec* [25]	**84.4**	83.0	**99.5**	65.6	83.13	4.1	40.8
LISTER* [8]	79.4	79.5	99.2	58.0	79.02	13.9	55.0
DPTR* [61]	80.0	79.6	98.9	64.4	80.73	0.0	68.0
CPPD* [13]	82.7	82.4	99.4	62.3	81.72	0.0	32.1
IGTR-AR* [14]	82.0	81.7	**99.5**	63.8	81.74	0.0	29.2
SMTR* [12]	83.4	83.0	99.3	65.1	82.68	49.4	20.8
CRNN* [39]	63.8	68.2	97.0	46.1	68.76	37.6	19.5
SVTR-B* [11]	77.9	78.7	99.2	62.1	79.49	22.9	19.8
SVTRv2-T	77.8	78.8	99.2	62.0	79.45	47.8	6.8
SVTRv2-S	81.1	81.2	99.3	65.0	81.64	50.0	14.0
SVTRv2-B	83.5	**83.3**	**99.5**	**67.0**	**83.31**	**52.8**	22.5

- Compared with previous methods, including EDTRs (CPPD, PARSeq, MAERec, and LISTER) and the CTC-based model (SVTR).
- SVTRv2 achieves new state-of-the-art performance in every scenario, including standard regular (Common) and irregular text (Curved and Multi-Oriented), Union14M-Benchmark, occluded scene text (OST), long text (Long), and Chinese text.
- Although SVTRv2 does not achieve the highest FPS, it remains the fastest among all EDTR models.