Horizontal/Curved Text Recognition

Model	IIIT	SVT	IC13	IC15	SVTP	CUTE
SemiMTR	97.3	96.6	97.0	84.7	93.0	93.8
MATRN	96.6	95.0	97.9	86.6	90.6	93.5

Horizontal/Curved Text Recognition

1. Multimodal Semi-Supervised Learning for Text Recognition (2022) (paper, code)

Table 2: **Scene text SOTA comparison.** Scene text recognition accuracies (%) over common and non-common public benchmarks. We show the number of words in each dataset below its title and present weighted (by size) average results on each set of datasets. The best performing result at each column is marked in bold. "*" refers to reproduced results and "Git" to GitHub model

			Common Benchmarks							Non-Common Benchmarks							
Method	**Labeled Data**	**Unlabeled Data**	**IIIT** 3,000	**SVT** 647	**IC13** 1,015	**IC15** 2,077	**SVTP** 645	**CUTE** 288	**Avg.** 7,672	**COCO** 9,835	**RCTW** 1,050	**Uber** 80,826	**ArT** 35,284	**LSVT** 4,257	**MLT19** 5,693	**ReCTS** 2,592	**Avg.** 139,537
PlugNet [28]	Synth	✗	94.4	92.3	95.0	82.2	84.3	85.0	89.8	-	-	-	-	-	-	-	-
RobustScanner [61]	Synth	✗	95.3	88.1	94.8	77.1	79.5	90.3	88.2	-	-	-	-	-	-	-	-
SCATTER [28]	Synth	✗	93.7	92.7	93.9	82.2	86.9	87.5	89.7	-	-	-	-	-	-	-	-
Plugnet [34]	Synth	✗	94.4	92.3	95.0	82.2	84.3	85.0	89.8	-	-	-	-	-	-	-	-
SRN [59]	Synth	✗	94.8	91.5	95.5	82.7	85.1	87.8	90.3	-	-	-	-	-	-	-	-
VisionLAN [56]	Synth	✗	95.8	91.7	95.7	83.7	86.0	88.5	91.1	-	-	-	-	-	-	-	-
TRBA [4]	Synth	✗	92.1	88.9	93.1	74.7	79.5	78.2	85.7	50.2	59.1	36.7	57.6	58.0	80.3	80.6	46.3
TRBA [5]	Real-L	✗	93.5	87.5	92.6	76.0	78.7	86.1	86.6	62.7	67.7	52.7	63.2	68.7	85.8	83.4	58.6
TRBA_{PL} [5]	Real-L	Real-U	94.8	91.3	94.0	80.6	82.7	88.1	89.3	66.9	71.5	54.2	66.7	73.5	87.8	85.6	60.9
TRBA_{PL} [5]	Real-L,Synth	Real-U	95.2	92.0	94.7	81.2	84.6	88.7	90.0	-	-	-	-	-	-	-	-
TRBA_{PL} [5]	Real-L,Synth	Real-U	95.2	92.0	94.7	81.2	84.6	88.7	90.0	-	-	-	-	-	-	-	-
$\text{ABINet}^{\text{Git}}$ [13]	Synth	✗	96.4	93.2	95.1	82.1	89.0	89.2	91.2	63.1	59.7	39.6	68.3	59.5	85.0	86.7	52.0
ABINet* [13]	Real-L	✗	95.5	93.4	94.4	83.0	87.1	89.6	90.8	69.2	71.6	55.7	67.7	73.7	88.2	90.6	62.4
ABINet_{PL}* [13,5]	Real-L	Real-U	96.4	94.1	95.0	83.7	88.8	93.1	91.8	71.2	74.2	56.8	70.5	75.0	89.1	90.9	63.9
ABINet_{est}* [13]	Real-L	Real-U	96.5	**96.3**	95.7	83.7	89.1	92.0	92.1	71.7	73.8	56.8	70.1	75.7	89.3	91.6	63.9
SemiMTR-V	Real-L	Real-U	95.6	93.5	95.2	82.5	88.1	90.6	91.0	70.5	75.1	57.7	69.5	75.2	89.6	92.3	64.2
SemiMTR-F	Real-L	Real-U	96.5	95.4	96.5	84.2	89.6	90.6	92.3	70.9	74.9	57.7	70.3	75.5	89.3	91.5	64.4
SemiMTR	Real-L	Real-U	**96.7**	95.5	**96.6**	**83.8**	**90.5**	**93.8**	**92.4**	**72.0**	**75.8**	**58.5**	**70.8**	**77.1**	**90.3**	**92.5**	**65.2**
SemiMTR	Real-L,Synth	Real-U	**97.3**	**96.6**	**97.0**	**84.7**	**93.0**	**93.8**	**93.3**	**72.7**	**76.3**	58.4	**72.3**	**77.1**	90.2	**93.2**	**65.6**

2. Multi-modal Text Recognition Networks: Interactive Enhancements between Visual and Semantic Features (2022) (paper, code)

Table 1. Recognition accuracies (%) on eight benchmark datasets, including the variant versions. The underlined values represent the best performances among the previous STR methods and the bold values indicate the best performances among all models including ours. For our implementation, we conduct repeated experiments with three different random seeds and report the averaged accuracy with standard deviation.

Model	Year	Regular test dataset				Irregular test dataset			
		IIIT	SVT	$IC13_S$	$IC13_L$	$IC15_S$	$IC15_L$	SVTP	CUTE
CombBest [2]	2019	87.9	87.5	93.6	92.3	77.6	71.8	79.2	74.0
ESIR [35]	2019	93.3	90.2	-	91.3	-	76.9	79.6	83.3
SE-ASTER [23]	2020	93.8	89.6	-	92.8	80.0		81.4	83.6
DAN [29]	2020	94.3	89.2	-	93.9	-	74.5	80.0	84.4
RobustScanner [34]	2020	95.3	88.1	-	94.8	-	77.1	79.5	90.3
AutoSTR [37]	2020	94.7	90.9	-	94.2	81.8	-	81.7	-
Yang *et al.* [32]	2020	94.7	88.9	-	93.2	79.5	77.1	80.9	85.4
SATRN [16]	2020	92.8	91.3	-	94.1	-	79.0	86.5	87.8
SRN [33]	2020	94.8	91.5	95.5	-	82.7	-	85.1	87.8
GA-SPIN [36]	2021	95.2	90.9	-	94.8	82.8	79.5	83.2	87.5
PREN2D [31]	2021	95.6	94.0	96.4	-	83.0	-	87.6	91.7
JVSR [3]	2021	95.2	92.2	-	95.5	-	**84.0**	85.7	89.7
VisionLAN [30]	2021	95.8	91.7	95.7	-	83.7	-	86.0	88.5
ABINet [7]	2021	96.2	93.5	97.4	-	86.0	-	89.3	89.2
ABINet (reproduced)		96.2 ±0.2	93.7 ±0.4	97.2 ±0.2	95.4 ±0.2	85.9 ±0.2	82.1 ±0.1	89.3 ±0.4	89.0 ±0.3
MATRN (ours)		**96.6** ±0.1	**95.0** ±0.2	**97.9** ±0.1	**95.8** ±0.1	**86.6** ±0.1	82.8 ±0.0	**90.6** ±0.2	**93.5** ±0.6

3. Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition (2021) (paper, code, demo)

Arbitrarily Oriented Text Recognition

1. SwinTextSpotter: Scene Text Spotting via Better Synergy between Text Detection and Text Recognition (2022) (paper, code)
2. Visual Semantics Allow for Textual Reasoning Better in Scene Text Recognition (2022) (paper, code)

Methods	Training Data	Regular			Irregular			Params ($\times 10^6$)	Time (ms)
		IIIT5k	SVT	IC13	SVTP	IC15	CUTE		
CRNN (Shi, Bai, and Yao 2016)	ST + MJ	78.2	80.9	86.7	-	-	-	8.3	6.8
ASTER (Shi et al. 2018)	ST + MJ	93.4	89.5	91.8	78.5	76.1	79.5	22	73.1
TRBA (Baek et al. 2019)	ST + MJ	87.9	87.5	92.3	79.2	77.6	74.0	49.6	27.6
Textscanner* (Wan et al. 2020a)	ST + MJ	93.9	90.1	92.9	84.3	79.4	83.3	57	56.8
GTC (Hu et al. 2020)	ST + MJ	95.5	92.9	94.3	86.2	82.5	92.3	-	-
SCATTER (Litman et al. 2020)	ST + MJ	93.7	92.7	93.9	86.9	82.2	87.5	-	-
SEED (Qiao et al. 2020)	ST + MJ	93.8	89.6	92.8	81.4	80.0	83.6	-	-
SRN (Yu et al. 2020)	ST + MJ	94.8	91.5	95.5	85.1	82.7	87.8	49.3	26.9
RobustScanner (Yue et al. 2020)	ST + MJ	95.3	88.1	94.8	79.5	77.1	90.3	-	-
Base2D (Yan et al. 2021)	ST + MJ	95.4	93.4	95.9	86.0	81.9	89.9	59.0	61.6
PREN2D (Yan et al. 2021)	ST + MJ	95.6	94.0	96.4	87.6	83.0	91.7	-	67.4
ABINet-LV† (Fang et al. 2021)	ST + MJ	96.3	93.0	97.0	88. 5	85.0	89.2	36.7	22.0
Seg-Baseline	ST + MJ	94.2	90.8	93.6	84.3	82.0	87.6	34.0	14.0
S-GTR	ST + MJ	**95.8**	**94.1**	**96.8**	**87.9**	**84.6**	**92.3**	42.1	18.8
GTR + CRNN$^{[CTC]}$	ST + MJ	87.6	82.1	90.1	68.1	68.2	78.1	15.2	12.8
GTR + TRBA$^{[1DATT]}$	ST + MJ	93.2	90.1	94.0	80.7	76.0	82.1	54.2	32.9
GTR + SRN$^{[Transformer]}$	ST + MJ	96.0	93.1	96.1	87.9	83.9	90.7	54.3	31.6
GTR + Base2D$^{[2DATT]}$	ST + MJ	96.1	94.1	96.6	88.0	85.3	92.6	64.1	65.7
GTR + ABINet-LV$^{\dagger[Transformer]}$	ST + MJ	96.8	94.8	97.7	89.6	86.9	93.1	41.6	30.9
SAR(Li et al. 2019)	ST + MJ + R	95.0	91.2	94.0	86.4	78.0	89.6	-	-
Textscanner* (Wan et al. 2020a)	ST + MJ + R	95.7	92.7	94.9	84.8	83.5	91.6	57	56.8
RobustScanner (Yue et al. 2020)	ST + MJ + R	95.4	89.3	94.1	82.9	79.2	92.4	-	-
ABINet (Fang et al. 2021)	ST + MJ + R	97.2	95.5	97.7	90.1	86.9	94.1	-	-
S-GTR	ST + MJ + R	**97.5**	**95.8**	**97.8**	**90.6**	**87.3**	**94.7**	42.1	18.8

Table 1: Results of our S-GTR, SOTA methods and their variants with our GTR on six regular and irregular STR datasets. "R" denotes the real datasets. "*" means using character-level annotations during training. "†" means the batch size is set to 384 for a fair comparison. The superscripts in the second group of rows denote the type of different methods, *i.e.*, "CTC": CTC-based method, "1DATT": 1D attention-based method, "2DATT": 2D attention-based method, and "Transformer": Transformer-based method. Details can be found in Section .

3. Text Spotting Transformers (2022) (paper, code)
4. A Bilingual, Open World Video Text Dataset and End-to-end Video Text Spotter with Transformer (2021) (paper, code)
5. Language Matters: A Weakly Supervised Vision-Language Pre-training Approach for Scene Text Detection and Spotting (2022) (paper, code)
6. Text Perceptron: Towards End-to-End Arbitrary-Shaped Text Spotting (2020) (paper, code)
7. TextAdaIN: Paying Attention to Shortcut Learning in Text Recognizers (2022) (paper, code)
8. TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models (2023) (paper, code)