Revisiting Scene Text Recognition: A Data Perspective

Qing Jiang , Jiapeng Wang , Dezhi Peng , Chongyu Liu , Lianwen Jin*
South China University of Technology
mountchicken@outlook.com , scutjpwang@foxmail.com, eedzpeng@mail.scut.edu.cn
liuchongyu1996@gmail.com , eelwjin@scut.edu.cn
https://union14m.github.io/

Abstract

This paper aims to re-assess scene text recognition (STR) from a data-oriented perspective. We begin by revisiting the six commonly used benchmarks in STR and observe a trend of performance saturation, whereby only 2.91% of the benchmark images cannot be accurately recognized by an ensemble of 13 representative models. While these results are impressive and suggest that STR could be considered solved, however, we argue that this is primarily due to the less challenging nature of the common benchmarks, thus concealing the underlying issues that STR faces. To this end, we consolidate a large-scale real STR dataset, namely Union14M, which comprises 4 million labeled images and 10 million unlabeled images, to assess the performance of STR models in more complex real-world scenarios. Our experiments demonstrate that the 13 models can only achieve an average accuracy of 66.53% on the 4 million labeled images, indicating that STR still faces numerous challenges in the real world. By analyzing the error patterns of the 13 models, we identify seven open challenges in STR and develop a challenge-driven benchmark consisting of eight distinct subsets to facilitate further progress in the field. Our exploration demonstrates that STR is far from being solved and leveraging data may be a promising solution. In this regard, we find that utilizing the 10 million unlabeled images through self-supervised pre-training can significantly improve the robustness of STR model in real-world scenarios and leads to state-of-the-art performance. Code and dataset is available at https://github.com/Mountchicken/Union14M.

1. Introduction

The success of deep learning in visual recognition tasks heavily depends on expansive labeled data. A widely used paradigm [2, 9, 10, 32, 62] in STR is training models on large-scale synthetic datasets [17, 18, 12, 61] and evaluating on six real benchmarks [43, 42, 55, 38, 21, 20]. Promisingly, current progress in STR has exhibited a trend of accuracy saturation (depicted in Fig. 1). The challenges in the common benchmarks seem "solved", suggested by the narrow scope for improvement, and the slowdown step of performance gain in recent SOTAs. This phenomenon inspires us to raise questions of 1) *whether the common benchmarks remain sufficient to promote future progress,* and 2) *whether this accuracy saturation implies that STR is solved.*

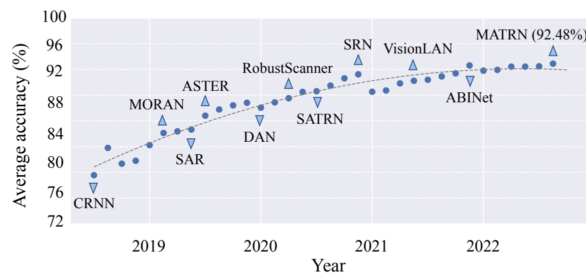

Figure 1. Average accuracy of STR models on six commonly used benchmarks as reported in their original papers. Models are trained with synthetic data.

For the first question, we start by selecting 13 representative models (listed in Tab. 3), including CTC-based [47, 9], attention-based [48, 34, 25, 24, 46, 56, 63], and language model-based [62, 10, 57, 39] models. We then evaluate their performance on the six STR benchmarks to find their joint errors. As depicted in Fig. 2, only 3.9% (298 images) of the total 7672 benchmark images can not be correctly recognized by any of the 13 models, among which 25.5% of the images are incorrectly annotated, and 35.2% images are barely recognizable (human unrecognizable samples). This suggests that there might be a maximum of 2.91% (222 images) and a minimum of 1.53% (117 images, excluding human unrecognizable samples) scope for accuracy improvement. Therefore, the common benchmarks give limited insight into future STR research.

The accuracy saturation in common benchmarks can obscure challenges that STR models still face. Therefore,

*Corresponding author.

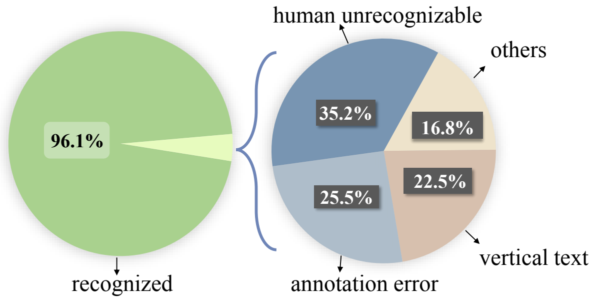

Figure 2. Error analysis on the six STR benchmarks.

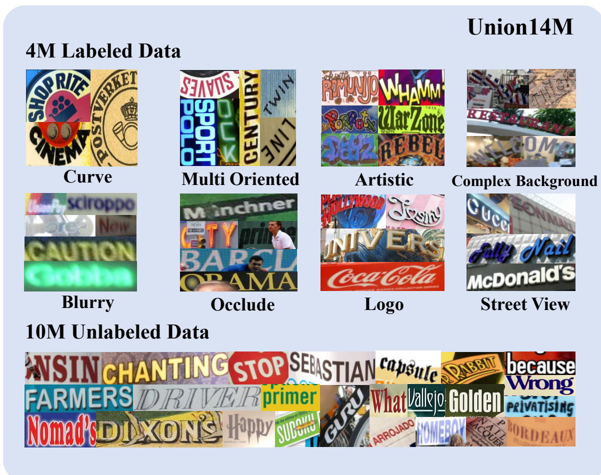

Figure 3. An Overview of Union14M which is used to analyze STR models in real-world scenarios. Union14M contains 4M labeled images (Union14M-L) and 10M unlabeled images (Union14M-U), which cover a wide range of real-world scenarios with intense diversity and complexity.

to bring more profound insights beyond these benchmarks and to benefit real-world STR applications, we consolidate a large-scale real dataset, namely Union14M (Fig. 3), to carefully analyze the performance of STR models in a broader range of real-world scenarios. Union14M consists of 4 million **L**abeled images (Union14M-L) and 10 million **U**nlabeled images (Union14M-U), obtained from 17 publicly available datasets. Hence, it can be considered as a comprehensive representation of text in the real world.

Equipped with Union14M-L, we conducted a quantitative evaluation on the aforementioned 13 STR models from multiple perspectives, thereby uncovering challenges that remain in STR. Our initial observation is related to the synthetic-to-real training paradigm. We discover that models trained on synthetic data perform poorly on Union14M-L, with an average accuracy of only 66.53%, despite achieving an average accuracy of 87.03% on commonly used benchmarks. This result indicates that such a paradigm is not compatible with more complex real-world settings. Subsequently, we analyze the error patterns of the 13 models and find that they are still less robust to four existing challenges, namely *curve text*, *multi-oriented text*, *contextless text*, and *artistic text*. Furthermore, we identify three additional challenges that are prevalent in the real world but have received less attention in the STR community, namely *multi-words text*, *salient text*, and *incomplete text*.

To enable more thorough evaluations of STR models in real-world scenarios and to encourage future research on the seven aforementioned challenges, we construct a challenge-driven benchmark, which comprises eight subsets with 400,000 generic samples and 9,383 challenge-specific samples sourced from Union14M-L. Extensive baseline experiments are conducted on this new benchmark and we find that despite utilizing real data for training, the current SOTA model can only achieve an average accuracy of 74.6%. This indicates that STR still faces numerous challenges in the real world and also answers the second question that STR is far from being solved.

Essentially, we infer that the sub-optimal performance of STR models in the real world can be attributed to data problems, e.g., the lack of sufficient real labeled data for training. To solve STR from a data perspective, we propose a solution of utilizing unlabeled data. Specifically, we investigate a Vision Transformer-based [8] STR model (Fig. 5), which can leverage the 10M unlabeled images in Union14M-U through self-supervised pre-training. The pre-trained ViT model exhibits powerful textual representation capabilities, and after fine-tuning on real labeled data, it achieves SOTA performance on both six common benchmarks and the proposed challenge-driven benchmark. Our contributions are summarized as follows:

- We analyze STR from a data perspective and arrive at two macro findings. Firstly, the common benchmarks are insufficient in presenting adequate challenges for advancing the field of STR. Secondly, despite significant progress, STR models still struggle to perform well in real-world scenarios. It is safe to say that STR is still far from being solved.
- We consolidate a large-scale real STR dataset to investigate the performance of STR models in the real world. Through quantitative analysis, we reveal that current STR models fail to address seven open challenges. Therefore, we propose a challenge-driven benchmark to facilitate future comprehensive and in-depth studies in the field of STR.
- We exploit the potential of unlabeled data and observe that they can lead to significant performance gains through self-supervised pre-training, offering a practical solution for STR in the real world.

2. Related Works

2.1. Data Analysis in STR

In scene text recognition, some works have been proposed to analyze several data issues. For instance, Baek *et al.* [2] point out the inconsistency between the training data and benchmarks in STR approaches. They also conduct a comprehensive analysis on the common benchmarks and find that 7.5% of the images can not be recognized by their proposed four-stage framework. In this work, we further refined it to 3.9% by using 13 distinctive STR models. Baek *et al.* [3] explored the impact of real data on the performance of STR models, in which they found that training on fewer real data can lead to better performance than training on synthetic data, and several recent works [60, 4, 50] have confirmed this finding by using real data for training. We also observe in our subsequent experiments that training models on real data can improve their generalization ability, which is essential for real-world STR applications.

2.2. Data Shift in STR

Scene text recognition is a fine-grained task that requires extensive amounts of training data. In the early time, due to the lack of sufficient real annotated data, STR models were trained on large-scale synthetic datasets, e.g., MJ [17, 18] and ST [12].This training paradigm still prevails today as state-of-the-art methods [10, 24, 39] continuously yield better performance on the common benchmarks. Nevertheless, models trained on synthetic data might suffer from generalization problems, due to the large domain gap [67, 3] between synthetic data and real-world circumstances.

Meanwhile, a few annotated real datasets have emerged in recent years [7, 22, 50, 41, 31]. Serval recent works have endeavored to consolidate these datasets. For instance, the OOV [11] dataset is a consolidation of seven real datasets and is employed to investigate the out-of-vocabulary [54] problem. Baek *et al.* [3], Yang *et al.* [60], and Darwin *et al.* [4] use different amounts of real datasets to construct the training set respectively, and achieved better results than training on synthetic data. In this work, our aim is to analyze the performance of STR models in the real world and the challenges they confront. Therefore, we consolidate Union14M with more real datasets, thus it can be used as a real-world mapping for our analysis.

2.3. Benchmarks in STR

In STR, there are six commonly used benchmarks, including regular text benchmarks: IC13 [21], IIIT [38], SVT [55] and irregular text benchmarks: IC15 [20], SVTP [42], CUTE [43]. Some recent works [60, 4] attempt to use alternative benchmarks [53, 29, 6, 7] for evaluation, and they also observe performance degradation on these benchmarks compared to the six benchmarks. This suggests that there exists challenges that exceed the scope of common benchmarks and an in-depth analysis is necessary.

Table 1. Composition of Union14M. † denotes that the dataset overlaps with current benchmarks. ‡ denotes those datasets overlap with each other. #Original denotes the number of text instances in the original dataset.

	Dataset	Year	#Original	#Refined	Lang.
Union14M-L	KAIST [19]	2011	6K	2K	EN, KR
	NEOCR [40]	2011	5K	3K	EN
	Uber-Text [66]	2017	209K	208K	EN
	RCTW [49]	2017	44K	7K	EN, CH
	IIIT-ILST [37]	2017	6K	2K	EN, IN
	MTWI [15]	2018	139K	53K	EN, CN
	COCOTextV2 [53]	2018	201K	73K	EN
	LSVT [51, 52]	2019	382K	38K	EN, CN
	MLT19 [41]	2019	89K	56K	Multi
	ReCTS [65]	2019	109K	25K	EN, CN
	ArT† [7]	2019	50K	35K	EN, CN
	IntelOCR‡ [22]	2021	2.57M	2.01M	EN
	TextOCR‡] [50]	2021	822K	586K	EN
	HierText‡ [31]	2022	1.2M	945K	EN
Union14M-U	Book32 [16]	2016	-	2.7M	-
	CC [45]	2018	-	5.6M	-
	OpenImages‡ [23]	2020	-	2.3M	-

3. Preliminary: A Real Dataset for Analysis

As previously discussed, the six STR benchmarks have almost reached a point of saturation, and can be insufficient to facilitate our analysis across a broader spectrum of real-world scenarios. Hence, we consolidate a large-scale real STR dataset denoted as Union14M, comprising 4 million labeled images (Union14M-L) and 10 million unlabeled images (Union14M-U), to support our subsequent analysis.

3.1. Dataset Consolidation

Union14M-L: 4M labeled images. Our data collection strategy is driven by the primary objective of encompassing a broad range of real scenarios. To this end, we collect labeled images from 14 publicly available datasets (Tab. 1) to compose Union14M-L. These datasets exhibit diverse properties. For instance, ArT [7] dataset is focused on curved text; ReCTS [65], RCTW [49], LSVT [51, 52], KAIST [19], NEOCR [40] and IIIT-ILST [37] datasets are designed for street views from different countries; MTWI [15] is sourced from web pages and contains scene text images; COCOTextV2 [53] contains plenty of low-resolution text images as well as vertical text images; IntelOCR [22], TextOCR [50] and HierText [31] are all derived from OpenImages [23], which is a vast dataset with nine million images covering an extensive range of real scenes. The consolidation of the 14 datasets can be viewed as a mapping of the real world, enabling our analysis to be oriented toward real-world scenarios.

Nevertheless, the simple concatenation of these 14 datasets is sub-optimal due to different annotation formats and the existence of duplicate, Non-Latin, and corrupted samples. Hence, we adopt the following strategies to refine.

- **Crop text instances.** Most datasets provide polygon annotations for text instances, and directly using the

Table 2. Statistics of Union14M and synthetic datasets MJ [17, 18] and ST [12]. Vertical instances are text images with a height that is at least twice their width and with more than one text character.

Dataset	# Instances	# Vocabularies	# Vertical Instances
MJ+ST	17M	384K	7K
Union14M-L	4M	707K	110K
Union14M-U	10M	-	39K

polygon for cropping is an intuitive choice. However, we conjecture this could be sub-optimal. Instead, we use the minimum axis-aligned rectangle for cropping, which can result in additional background noise for cropped text instances. This cropping strategy essentially serves as a form of regularization, as it introduces challenging samples (i.e., those with more background noise) that enhance the robustness of the recognizer. This is beneficial in an end-to-end system, as the recognizer can be less dependent on the performance of the detector, and also allows us to focus our analysis on the performance of the recognizer.

- **Exclude duplicate samples.** We first remove duplicate samples between Union14M-L and the common benchmarks. Next, we remove duplicate samples among the 14 datasets. For instance, HierText, TextOCR, and IntelOCR are duplicated with each other since they are all annotated from OpenImages [23]. We choose HierText as reference, and remove duplicated samples from the remaining two datasets.
- **Remove Non-Latin and ignored samples.** In this work, we focus on Latin characters which are widely employed and possess a large amount of data. Consequently, We only retained samples composed of letters, numbers, and symbols. We also remove samples that are annotated as ignored.

Union14M-U: 10M unlabeled images. Self-supervised learning has enabled substantial development in computer vision [59, 13, 14, 5], and several related works have also emerged in the field of STR [33, 60, 36, 1]. The optimal solution to improve the performance of STR in real-world scenarios is to utilize more data for training. However, labeling text images is both costly and time-intensive, given that it involves annotating sequences and needs specialized language expertise. Therefore, it would be desirable to investigate the potential of utilizing unlabeled data via self-supervised learning for STR. To this end, we collect 10M unlabeled images from three large datasets, including Book32 [16], OpenImages [23] and Conceptual Captions (CC) [45] dataset. To obtain high-quality text instances, we adopt a different collection method than previous works [60, 3]. We use three text detectors [68, 27, 30] and an IoU voting mechanism to get text instances (detailed in Appendix). The unlabeled images collected from OpenImages are also de-duplicated with the labeled images in Union14M-L.

Table 3. We use 13 publicly available models for evaluation. Acc-CB represents the average accuracy on six commonly used benchmarks. Acc-UL represents the accuracy on all Union14M-L data.

Method	Type	Venue	Acc-CB	Acc-UL
CRNN [47]	CTC	TPAMI'17	78.14	57.96 (-20.18)
SVTR [9]	CTC	IJCAI'22	90.00	69.46 (-20.54)
MORAN [34]	Att.	PR'17	80.61	57.73 (-22.88)
ASTER [48]	Att.	TPAMI'19	84.98	63.30 (-21.68)
NRTR [46]	Att.	ICDAR'19	86.82	66.96 (-19.86)
SAR [25]	Att.	AAAI'19	88.07	68.07 (-20.00)
DAN [56]	Att.	AAAI'20	83.96	64.16 (-19.80)
SATRN [24]	Att.	CVPRW'20	91.36	72.09 (-19.27)
RobustScanner [63]	Att.	ECCV'20	87.63	67.63 (-20.00)
SRN [62]	LM	CVPR'20	86.51	65.71 (-20.80)
ABINet [10]	LM	CVPR'21	91.97	70.73 (-21.24)
VisionLAN [57]	LM	ICCV'21	88.96	69.60 (-19.36)
MATRN [39]	LM	ECCV'22	92.48	71.49 (-20.99)

3.2. Characteristics of Real-World Data

Diverse text styles. As shown in Fig. 3, Union14M covers text images from a variety of real scenes. Real-world text images exhibit diverse layouts, e.g., curve, tilted and vertical, as well as challenging distractions, including blurring, complex background and occlusion, and also various real-world applications of scene text, such as street scenes and logos. Notably, Union14M contains a large number of vertical text instances (last column in Tab. 2), which are common in real world, yet are rare in synthetic datasets.

Large vocabularies. Text used in synthetic datasets are obtained from commonly used corpus. However, in real-world scenarios, there are plenty of text variations that are not covered by corpus, such as random combinations of alphanumeric characters and symbols, for instance, license plates, or multilingual alphabetical combinations like Chinese Pinyin. In Tab. 2, we show that the number of vocabularies in Union14M-L is nearly twice as larger as that of synthetic datasets, demonstrating that Union14M-L can encompass a broader spectrum of real-world situations and thus can hold our further analysis.

4. Analysis of STR in Real World

In this section, we utilize the vast nature of Union14M-L to conduct a comprehensive analysis of the performance of 13 STR models. The objective of this analysis is to evaluate the robustness of STR models against numerous real-world challenges, identify existing challenges, and stimulate future research advances.

4.1. Overall Performance Evaluation

We begin by selecting 13 representative models trained on synthetic datasets to evaluate on Union14M-L. As shown in Tab. 3, compare to the performance on common benchmarks, their performance degradation on Union14M-L is significant, with an average accuracy drop of 20.50%. This result suggests that models trained on synthetic data can

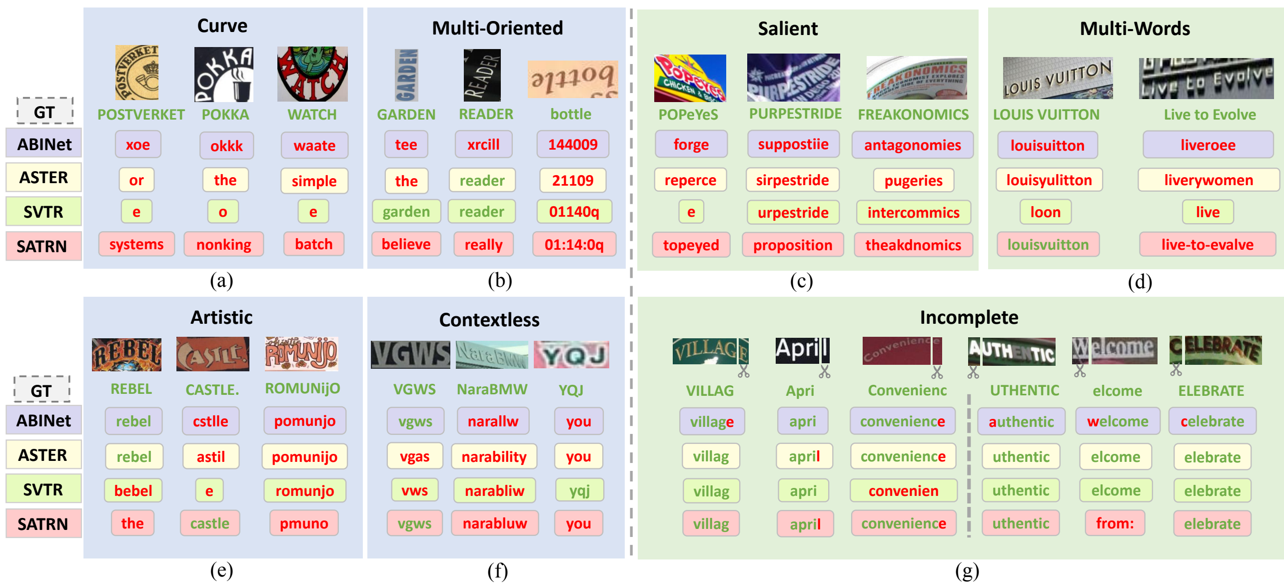

Figure 4. Error analysis of the 13 STR models. We select four representative models and show their prediction results (Text in green represents correct prediction and red text vice versa). Blocks in blue (a, b, e, f) represent four unsolved challenges, and blocks in green (c, d, g) represent three additional challenges that are rarely discussed. Best viewed in color.

not well generalize to more complicated real-world scenarios. Conversely, it also suggests that Union14M-L features challenges that are not covered by common benchmarks and worth a deeper investigation.

4.2. Challenge Mining

To identify the joint errors made by the 13 models, we assign each sample in Union14M-L with a difficulty score based on the number of correct predictions (detailed in Appendix). We focus on the hard samples that the majority of models fail to make correct predictions, and we summarize four challenges that haven't been adequately solved (left side of Fig. 4). Furthermore, we introduce three additional challenges that are common in the real world, yet are seldom discussed in previous works (right side of Fig. 4).

4.2.1 Unsolved Challenges

Curve Text. Curve text recognition has gained considerable attention in recent years, with two mainstream approaches: one that relies on rectification [34, 48] and the other that employs 2D attention mechanism [24, 25, 63]. Both approaches yield promising results on the curve text benchmark CUTE [43]. However, the proportion of curve text in the CUTE benchmark is limited, and the extent of curvature is minor. For highly curved texts as shown in Fig. 4a, current methods still exhibit limited performance.

Multi-Oriented Text. Text can appear on the surface of any object in any orientation, including vertical, tilted, or reversed cases (Fig. 4b). Multi-oriented text is common in real-world scenarios, such as vertical text on billboards, tilted text due to the shooting angle of the camera, and reversed text due to mirror reflection. However, this problem is overlooked in most STR methods with an assumption that text images are nearly horizontal. They followed a similar procedure of scaling the height of text images to a small size (e.g., 32 pixels), and then scaling the width while keeping the ratio unchanged, causing vertical or tilted images to collapse in height and consequently impeding recognition.

Artistic Text. In contrast to printed text, artistic text is designed by artists or professional designers with diverse text fonts, text effects, text layouts, and complex backgrounds. Each instance of the artistic text is potentially unique, making it a zero-shot or one-shot problem, and may require specifically designed networks [58] for recognition. Nevertheless, due to the lack of artistic text samples in the synthetic datasets, current models are still less robust to the artistic text shown in Fig. 4e.

Contextless Text. Contextless text refers to text that has no semantic meaning and is not in the dictionary. It can be abbreviations or random combinations of letters, digits, and symbols. As shown in Fig. 4f, models may fail to recognize contextless text even when it has a clear background and minimal distortion. This issue can arise from the over-introduction of semantic information in both the model design and dataset corpus, which is also known as vocabulary reliance [54, 11]. Models will attempt to predict text that appeared in the training set that follows syntax rules (e.g., mistaking "YQJ" for "you" in Fig. 4f). This behavior is highly undesirable in applications where reliability is critical, e.g., license plate recognition, invoice recognition, and card ID recognition, where most of the text are contextless and their misrecognition can lead to enormous security risks and property damages.

4.2.2 Additional Challenges

Salient Text. Salient text refers to the presence of extra characters that coexist with the primary characters of interest in a text image (Fig. 4c). Salient text can be inadvertently introduced in end-to-end text recognition when text instances of different sizes are adjacent or overlapping with each other. This problem has been discussed in the text detection stage. For instance, Liao *et al.* [26] propose to use a hard ROI masking strategy to eliminate the interference of extra characters. Nevertheless, when the performance of the detection model is poor, e.g., when it can only output coarse text regions, it becomes crucial for recognition models to rapidly identify visually important regions. However, as shown in Fig. 4c, models can be confused by additional characters and fail to recognize the primary text.

Multi-Words Text. Text contains rich semantic information that aids in the comprehension of scenes, and sometimes a single word may be insufficient. In certain cases, the recognition of multiple words simultaneously is required to fully interpret a text image, such as trademarks and short phrases, as depicted in Fig. 4d. However, most STR models are trained on synthetic datasets that comprise a single word per text image, hence failing to recognize spaces that separate individual words. Moreover, We observe that models tend to amalgamate multiple words into a single word, discarding or altering visible characters based on syntax rules (e.g., "Live to Evolve" being identified as "liveroee" as it reads more like a single word)."

Incomplete Text. Text images can be incomplete, with missing characters due to occlusion or inaccurate detection boxes that truncate the text. In Fig. 4g, when a text image is cropped with the first or the last letter, models may produce completed predictions, even though the missing letter is invisible. Moreover, we observe that this behavior occurs more frequently in language models (Sec. 6.2) that rely heavily on linguistic priors. This feature may reduce the reliability of models in text analysis applications. For instance, a fragmented text image with "ight" written on it may be completed as "might" or "light", while it would be optimal for the recognition model to output what it actually sees, i.e. "ight", thus allowing anomaly detection. Therefore, it is crucial to thoroughly evaluate the performance of the automatic completion feature and consider the potential impact on downstream applications.

5. A Challenge-Driven Benchmark

To facilitate the evaluation of STR models in more comprehensive real-world scenarios and to support future research on the aforementioned seven challenges, we construct a challenge-driven benchmark, namely Union14M-Benchmark. It consists of eight subsets and a total of 409,393 images with both complexity and versatility.

Table 4. Dataset partion of Union14M-L. Union14M-Benchmark is a split from Union14M-L.

Dataset	#Images		
	Train	Val	Benchmark
Union14M-L	3,230,742	400,000	409,383
General	-	-	400,000
Artistic	-	-	900
Curve	-	-	2426
Multi-Orinted	-	-	1369
Multi-Words	-	-	829
Salient	-	-	1585
Incomplete	-	-	1495
Contextless	-	-	779

5.1. Benchmark Construction

Challenge-specific subsets. We collected subsets for each of the seven challenges presented in Sec. 4.2. Candidate images are manually selected from Union14M-L based on some reference samples of these seven text types, except for the incomplete text. For the incomplete text subset, we sample 1,495 images that the majority of the 13 models can make correct predictions from Union14M-L since we aim to investigate the auto-completion feature of STR models and therefore we shall not introduce other factors that might lead to false recognition. Then we randomly crop out either the first or the last letter of the text image. To ensure that there are no duplicate images between Union14M-L and the proposed benchmark, we counted the remaining samples in Union14M-L that have the same text label as the benchmark images, and then we manually reviewed each sample to remove the duplicate images in Union14M-L.

General subset. In addition to these seven specific challenges, STR poses several other difficulties, such as blurring, chromatic distortion [64], and complex background [35]. Therefore, to enhance the diversity of this benchmark, we also construct a general subset with 400,000 images sampled from Union14M-L.

We also emphasize the significance of the validation set. It follows the same construction methodology as the general subset, which also includes 400,000 samples. The statistics are shown in Tab. 4.

6. Experiments and Analysis

In this section, we benchmark the aforementioned 13 STR models (Tab. 3) on Union14M-L to provide more quantitative analysis. In addition, we also introduce a solution for STR from a data perspective by proposing a ViT-based model [8], namely **MAERec** (Sec. 6.4), which can utilize the 10M unlabeled images in Union14M-U through self-supervised pre-training.

Table 5. Performance (WAICS) of models trained on **synthetic datasets** (MJ and ST). For the incomplete text subset, we measure the margin of accuracy before and after image cropping, which is the lower the better.

Type	Method	Common Benchmarks							Union14M-Benchmark								
		IIIT 3000	IC13 1015	SVT 647	IC15 2077	SVTP 645	CUTE 288	Avg	Curve	Multi-Oriented	Artistic	Contextless	Salient	Multi-Words	General	Avg	Incomplete ↓
CTC	CRNN [47]	89.7	88.3	82.2	69.3	67.8	71.2	78.1	7.5	0.9	20.7	25.6	13.9	25.6	32.0	18.0	6.4
	SVTR [9]	94.4	96.3	91.6	**84.1**	85.4	88.2	90.0	63.0	**32.1**	37.9	44.2	67.5	49.1	52.8	**49.5**	4.8
Attention	MORAN [34]	91.0	91.3	83.9	68.4	73.3	75.7	80.6	8.9	0.7	29.4	20.7	17.9	23.8	35.2	19.5	6.8
	ASTER [48]	93.3	90.8	90.0	74.7	80.2	80.9	85.0	34.0	10.2	27.7	33.0	48.2	27.6	39.8	31.5	5.8
	NRTR [46]	95.2	94.0	90.0	74.1	79.4	88.2	86.8	31.7	4.4	36.6	37.3	30.6	**54.9**	48.0	34.8	7.3
	SAR [25]	95.0	93.7	89.6	79.0	82.2	88.9	88.1	44.3	7.7	42.6	44.2	44.0	51.2	50.5	40.6	**4.5**
	DAN [56]	93.4	92.1	87.5	71.6	78.0	81.3	84.0	26.7	1.5	35.0	40.3	36.5	42.2	42.1	37.4	6.7
	SATRN [24]	96.1	95.7	93.5	**84.1**	88.5	90.3	91.4	51.1	15.8	**48.0**	45.3	62.7	52.5	58.5	47.7	5.6
	RobustScanner [63]	95.1	93.1	89.2	77.8	80.3	90.3	87.6	43.6	7.9	41.2	42.6	44.9	46.9	39.5	38.1	**4.5**
LM	SRN [62]	91.5	93.9	88.9	76.0	84.0	84.8	86.5	**63.4**	25.3	34.1	28.7	56.5	26.7	46.3	39.6	7.6
	ABINet [10]	95.7	95.7	94.6	85.1	90.4	90.3	92.0	59.5	12.7	43.3	38.3	62.0	50.8	55.6	46.0	17.9
	VisionLAN [57]	95.9	94.4	90.7	80.1	85.3	88.9	89.2	57.7	14.2	47.8	**48.0**	64.0	47.9	52.1	47.4	6.9
	MATRN [39]	**96.7**	**95.8**	**94.9**	82.9	**90.5**	**94.1**	**92.5**	63.1	13.4	43.8	41.9	**66.4**	53.2	**57.0**	48.4	8.2

Table 6. Performance (WAICS) of models trained on the training set of **Union14M-L**. For MAERec, S and B represent the use of ViT-Small and ViT-Base as the backbone, respectively. PT denotes pre-training.

Type	Method	Common Benchmarks							Union14M-Benchmark								
		IIIT 3000	IC13 1015	SVT 647	IC15 2077	SVTP 645	CUTE 288	Avg	Curve	Multi-Oriented	Artistic	Contextless	Salient	Multi-Words	General	Avg	Incomplete ↓
CTC	CRNN [47]	90.8	91.8	83.8	71.8	70.4	80.9	81.6	19.4	4.5	34.2	44.0	16.7	35.7	60.4	30.7	**0.9**
	SVTR [9]	95.9	95.5	92.4	83.9	85.7	93.1	91.1	72.4	68.2	54.1	68.0	71.4	67.7	77.0	68.4	2.0
Attention	MORAN [34]	94.7	94.3	89.0	78.8	83.4	87.2	87.9	43.8	12.8	47.3	55.1	45.7	54.6	44.7	43.4	1.9
	ASTER [48]	94.3	92.6	88.9	77.7	80.5	86.5	86.7	38.4	13.0	41.8	52.9	31.9	49.8	66.7	42.1	1.3
	NRTR [46]	96.2	96.9	94.0	80.9	84.8	92.0	90.8	49.3	40.6	54.3	69.6	42.9	75.5	75.2	58.2	1.5
	SAR [25]	96.6	96.0	92.4	82.0	85.7	92.7	90.9	68.9	56.9	60.6	73.3	60.1	74.6	76.0	67.2	2.1
	DAN [24]	95.5	95.2	88.6	78.3	79.9	86.1	87.3	46.0	22.8	49.3	61.6	44.6	61.2	67.0	50.4	2.3
	SATRN [56]	97.0	97.9	95.2	87.1	91.0	96.2	93.9	74.8	64.7	67.1	76.1	72.2	74.1	75.8	72.1	0.9
	RobustScanner [63]	96.8	95.7	92.4	86.4	83.9	93.8	91.2	66.2	54.2	61.4	72.7	60.1	74.2	75.7	66.4	1.9
LM	SRN [62]	95.5	94.7	89.5	79.1	83.9	91.3	89.0	49.7	20.0	50.7	61.0	43.9	51.5	62.7	48.5	2.2
	ABINet [10]	97.2	97.2	95.7	87.6	92.1	94.4	94.0	75.0	61.5	65.3	71.1	72.9	59.1	79.4	69.2	2.6
	VisionLAN [57]	96.3	95.1	91.3	83.6	85.4	92.4	91.3	70.7	57.2	56.7	63.8	67.6	47.3	74.2	62.5	1.3
	MATRN [39]	98.2	97.9	96.9	88.2	94.1	97.9	95.5	80.5	64.7	71.1	74.8	79.4	67.6	77.9	74.6	1.7
Ours	MAERec-S w/o PT	97.4	97.3	95.7	86.7	91.0	96.2	94.1	75.4	66.5	66.0	76.1	72.6	77.0	80.8	73.5	3.5
	MAERec-S	98.0	97.6	96.8	87.1	93.2	97.9	95.1	81.4	71.4	72.0	82.0	78.5	82.4	82.5	78.6	2.7
	MAERec-B w/o PT	97.3	97.8	96.6	87.1	92.6	95.8	94.5	76.5	67.5	65.7	75.5	74.6	77.7	81.8	74.2	3.2
	MAERec-B	**98.5**	**98.1**	**97.8**	**89.5**	**94.4**	**98.6**	**96.2**	**88.8**	**83.9**	**80.0**	**85.5**	**84.9**	**87.5**	**85.8**	**85.2**	2.6

6.1. Experiment Settings

Training settings. For the 13 STR models, we use their default hyper-parameters described in the original papers for a fair comparison, except that the number of the predicted character classes is unified to 91 (including digits, upper and lower case letters, symbols, and space).

Metrics. We use three evaluation metrics: word accuracy (WA), word accuracy ignoring case (WAIC) and word accuracy ignoring case and symbols (WAICS, most commonly used). For the incomplete text subset, we measure the margin of accuracy before and after the letter cropping.

6.2. Experiment Results

Real-world data is challenging. As shown in Tab. 5 and Tab. 6, compared to the performance on common benchmarks, models exhibit an average accuracy degradation of 48.5% and 33.0% on Union14M-Benchmark, when trained on synthetic datasets and Union14M-L respectively. This indicates that the text images in real-world scenarios is far more complex than the six commonly used benchmarks.

Real-world data is effective. Models trained on Union14M-L can gain an average accuracy improvement of 3.9% and 19.6% on common benchmarks and Union14M-Benchmark, respectively. The large performance boost on Union14M-Benchmark suggests that synthetic training data can hardly accommodate complex real-world demands, while using real data for training can largely overcome this generalization problem. Additionally, the relatively small performance gains on common benchmarks also imply their saturation.

STR is far from being solved. When trained only on Union14M-L, we observe that the maximum average accuracy on Union14M-Benchmark (excluding incomplete text subset) is only 74.6% (by MATRN [39] in Tab .6). This indicates that STR is far from being solved. Although relying on large-scale real data can bring a certain performance improvement, future efforts are still needed.

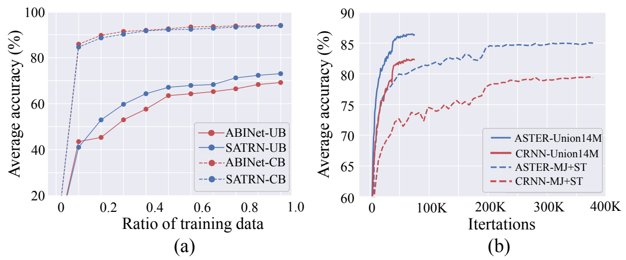

Figure 5. (a) Performance of models trained on increasing fractions of Union14M-L. CB denotes the six common benchmarks; UB denotes Union14M-Benchmark. (b) Performance evolution curves of models trained with Union14M-L or MJ [18, 17] and ST [12] under the same configurations (number of epochs, optimizer, etc.), evaluated on the six common benchmarks.

Vocabulary reliance is ubiquitous. When trained on synthetic datasets, all models exhibit a large performance drop on incomplete text subset. In particular, we observe that language models have a larger performance degradation (10.2% *vs.* 5.6% in CTC-based and 5.9% in attention-based models). We speculate that the performance drop in language models can be related to their error correction behavior, i.e., models complete the incomplete text which is viewed as a character missing error. This problem can be significantly alleviated when trained on Union14M-L. We attribute this to the larger vocabulary size in Union14M-L that models will not overfit the training corpus. However, this problem still exists and requires further investigation.

6.3. Data Saturation

We conducted a data ablation study to demonstrate the sufficiency of data in Union14M-L. We select ABINet[10] and SATRN[24], and train them on the increasing fractions of the Union14M-L dataset. As depicted in Fig. 5a, the accuracy increases sharply in the beginning and eventually levels out. This indicates that the data in Union14M-L are sufficient, and adding more real data may not lead to significant performance gain. Moreover, as shown in Fig. 5b, even though the data in Union14M-L are only 1/4 of the synthetic data, training on Union14M-L requires much fewer iterations (four times less) to achieve higher accuracy, which aligns with the Green AI[44] philosophy.

6.4. Exploration of Unlabeled Data

To further explore the potential of leveraging self-supervised pre-training to solve STR from a data perspective, we introduce a ViT-based model, namely MAERec.

Architecture of MAERec. In Fig. 6, we show the brief architecture of MAERec. We choose Vision Transformer (ViT) [8] as the default backbone for its effortless applicability in masked image modeling [13]. The input image is first fed into the ViT backbone with a patch size of 4×4. The output sequence is then passed to a Transformer decoder used in SATRN [24]. Specifically, we use six layers of Transformer decoder to predict text sequence in an auto-regressive manner. The embedding dimension of the Transformer decoder is set to 384 and 768 for the small and base models respectively. The number of heads is set to 8.

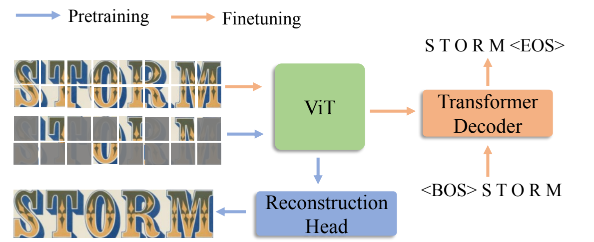

Figure 6. An overview of MAERec. It consists of a ViT [8] as the backbone and an auto-regressive Transformer decoder [24].

Figure 7. Reconstruction results on *Union14M-L* images (images that are not used during pre-training). For each triplet, we show the ground truth (top), the masked image (middle), and the reconstructed image (bottom). The mask ratio is 75%.

Pre-training. To utilize the 10M unlabeled images in Union14M-U, we pre-train the ViT backbone in MAERec through a masked image modeling task. We adopt the framework of MAE [13] with minor modifications. We use 75% mask ratio as in MAE. The mask size is set to 4×4 to be consistent with the patch size. We use the normalized pixel value of the input image as the reconstruction target since we find this leads to better results. The reconstruction results are shown in Fig. 7. The ViT backbone pre-trained on Union14-U can yield convincing reconstructed text images, despite the mask ratio being high up to 75%. This indicates that the pre-trained ViT backbone can effectively capture the text structure in text images and can learn useful textual representations. Notably, the ViT backbone pre-trained on Union14M-U can reconstruct images that have not been seen during pre-training, indicating that Union14M-U covers a wide range of real-world examples.

Fine-tuning. After pre-training, we initialize MAERec with the pre-trained ViT weight and fine-tune the whole model on Union14M-L. The results are shown in Tab. 6 (last four rows). The performance of MAERec can be substantially improved after pre-training, with an average accuracy gain of 1.0% on common benchmarks and 5.1% on the Union14M-Benchmark, when using ViT-Small as the backbone. Moreover, when scaling the backbone to ViT-Base, we can observe significant performance improve-

Table 7. Comparison between MAERec and other self-supervised learning-based STR models with different pre-training and fine-tuning data. R stands for real data, and S stands for synthetic data. We report the average accuracy on six common benchmarks.

Method	Pre-train	Fine-tune	Avg.
PerSec [28]	100M R	17M S	82.2
MaskOCR [36]	4.2M R, 100M S	17M S	92.6
DiG-S [60]	15.8M R, 17M S	2.8M R	94.6
DiG-B [60]	15.8M R, 17M S	2.8M R	95.0
MAERec-S (ours)	10.6M R	2.8M R	94.6
MAERec-S (ours)	10.6M R	3.2M R	95.1
MAERec-B (ours)	10.6M R	3.2M R	**96.2**

Table 8. Compare the performance of MAERec-S with different pre-training and fine-tuning datasets. Acc-CB denotes the average accuracy on six common benchmarks. Acc-UB denotes the average accuracy on Union14M-Benchmark (Exclude incomplete text subset).

No.	Pre-train	Fine-tune	Acc-CB	Acc-UB
1	-	MJ, ST	89.9	46.0
2	-	Union14M-L	94.1	73.5
3	MJ, ST	MJ, ST	89.9	46.1
4	MJ, ST	Union14M-L	94.0	75.0
5	Union14M-U	Union14M-L	**95.1**	**78.6**

ments and MAERec can achieve an average accuracy of 85.2% on Union14M-Benchmark. This promising result demonstrates that utilizing massive unlabeled data can substantially improve the performance of STR models in real-world scenarios, and it is worth further exploration.

Comparison with SOTA SSL methods. We also compare our proposed MAERec with other self-supervised learning-based methods in STR, as shown in Tab. 7. Despite the varying amounts of data used by different methods, MAERec outperforms its counterparts with a smaller data scale. It is noteworthy that MAERec utilizes a similar fine-tuning architecture with DiG [60] and a simpler pre-training framework, yet still achieves better results. This implies that the selection of data plays an even more critical role in self-supervised pre-training and fine-tuning.

Data Matters in Self-Supervised Pre-training. In Tab. 8, we compare different dataset combinations used in pre-training and fine-tuning. When pre-training and fine-tuning are both performed on synthetic datasets, MAERec can barely gain a performance boost (89.9% $\rightarrow$ 89.9% for CB, 46.0% $\rightarrow$ 46.1% for UB). However, when fine-tuning is performed on Union14M-L, MAERec can exhibit a performance boost when either pre-trained on synthetic datasets (73.5% $\rightarrow$ 75.0% for UB) or on Union14M-U (73.5% $\rightarrow$ 78.6% for UB). This indicates that fine-tuning on real data is vital for self-supervised learning, and Union14M-U is preferable to synthetic datasets for pre-training (78.6% *vs.* 75.0%).

7. Conclusion

In this paper, we revisit scene text recognition from a data perspective. Despite the current benchmarks being close to saturation, we argue that the problem of STR remains unsolved, especially in real-world scenarios where current models struggle with numerous challenges. To explore the challenges that STR models still face, we consolidate a large-scale STR dataset for analysis and identified seven open challenges. Furthermore, we propose a challenge-driven benchmark to facilitate the future development of STR. Additionally, we reveal that the utilization of massive unlabeled data through self-supervised pre-training can remarkably enhance the performance of the STR model in real-world scenarios, suggesting a practical solution for STR from a data perspective. We hope this work can spark future research beyond the realm of existing data paradigms.

Acknowledgments

This research is supported in part by NSFC (Grant No.: 61936003), National Key Research and Development Program of China (2022YFC3301703), Zhuhai Industry Core and Key Technology Research Project (no. 2220004002350).

References

[1] Aviad Aberdam, Ron Litman, Shahar Tsiper, Oron Anschel, Ron Slossberg, Shai Mazor, R Manmatha, and Pietro Perona. Sequence-to-sequence contrastive learning for text recognition. In *CVPR*, pages 15302–15312, 2021. 4

[2] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *ICCV*, pages 4715–4723, 2019. 1, 3

[3] 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 *CVPR*, pages 3113–3122, 2021. 3, 4

[4] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *ECCV*, pages 178–196. Springer, 2022. 3

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020. 4

[6] Chee Kheng Ch'ng and Chee Seng Chan. Total-text: A comprehensive dataset for scene text detection and recognition. In *ICDAR*, volume 1, pages 935–942. IEEE, 2017. 3

[7] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. ICDAR 2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *ICDAR*, pages 1571–1576. IEEE, 2019. 3

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 6, 8

[9] Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, and Yu-Gang Jiang. SVTR: Scene text recognition with a single visual model. *IJCAI*, pages 884–890, 2022. 1, 4, 7

[10] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. Read Like Humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *CVPR*, pages 7098–7107, 2021. 1, 3, 4, 7, 8

[11] Sergi Garcia-Bordils, Andrés Mafla, Ali Furkan Biten, Oren Nuriel, Aviad Aberdam, Shai Mazor, Ron Litman, and Dimosthenis Karatzas. Out-of-vocabulary challenge report. In *ECCV Workshops*, pages 359–375. Springer, 2023. 3, 5

[12] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *CVPR*, pages 2315–2324, 2016. 1, 3, 4, 8

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 4, 8

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 4

[15] Mengchao He, Yuliang Liu, Zhibo Yang, Sheng Zhang, Canjie Luo, Feiyu Gao, Qi Zheng, Yongpan Wang, Xin Zhang, and Lianwen Jin. ICPR 2018 contest on robust reading for multi-type web images. In *ICPR*, pages 7–12. IEEE, 2018. 3

[16] Brian Kenji Iwana, Syed Tahseen Raza Rizvi, Sheraz Ahmed, Andreas Dengel, and Seiichi Uchida. Judging a book by its cover. *arXiv preprint arXiv:1610.09204*, 2016. 3, 4

[17] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. In *NIPS Deep Learning Workshop*, 2014. 1, 3, 4, 8

[18] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *Int. J. Comput. Vis.*, 116(1):1–20, 2016. 1, 3, 4, 8

[19] Jehyun Jung, SeongHun Lee, Min Su Cho, and Jin Hyung Kim. Touch TT: Scene text extractor using touchscreen interface. *ETRI Journal*, 33(1):78–88, 2011. 3

[20] 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 *ICDAR*, pages 1156–1160. IEEE, 2015. 1, 3

[21] 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 *ICDAR*, pages 1484–1493. IEEE, 2013. 1, 3

[22] Ilya Krylov, Sergei Nosov, and Vladislav Sovrasov. Open images v5 text annotation and yet another mask text spotter. In *ACML*, pages 379–389. PMLR, 2021. 3

[23] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *Int. J. Comput. Vis.*, 128(7):1956–1981, 2020. 3, 4

[24] Junyeop Lee, Sungrae Park, Jeonghun Baek, Seong Joon Oh, Seonghyeon Kim, and Hwalsuk Lee. On recognizing texts of arbitrary shapes with 2d self-attention. In *CVPR Workshops*, pages 546–547, 2020. 1, 3, 4, 5, 7, 8

[25] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show, Attend and Read: A simple and strong baseline for irregular text recognition. In *AAAI*, volume 33, pages 8610–8617, 2019. 1, 4, 5, 7

[26] Minghui Liao, Guan Pang, Jing Huang, Tal Hassner, and Xiang Bai. Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In *ECCV*, pages 706–722. Springer, 2020. 6

[27] Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 4

[28] Hao Liu, Bin Wang, Zhimin Bao, Mobai Xue, Sheng Kang, Deqiang Jiang, Yinsong Liu, and Bo Ren. Perceiving stroke-semantic context: Hierarchical contrastive learning for robust scene text recognition. AAAI, 2022. 9

[29] Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Sheng Zhang. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognition*, 90:337–345, 2019. 3

[30] Yuliang Liu, Sheng Zhang, Lianwen Jin, Lele Xie, Yaqiang Wu, and Zhepeng Wang. Omnidirectional scene text detection with sequential-free box discretization. In *IJCAI*, pages 3052–3058, 2019. 4

[31] Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Towards end-to-end unified scene text detection and layout analysis. In *CVPR*, pages 1049–1059, 2022. 3

[32] Ning Lu, Wenwen Yu, Xianbiao Qi, Yihao Chen, Ping Gong, Rong Xiao, and Xiang Bai. MASTER: Multi-aspect non-local network for scene text recognition. *Pattern Recognition*, 117:107980, 2021. 1

[33] Canjie Luo, Lianwen Jin, and Jingdong Chen. SimAN: Exploring self-supervised representation learning of scene text via similarity-aware normalization. In *CVPR*, pages 1039–1048, 2022. 4

[34] Canjie Luo, Lianwen Jin, and Zenghui Sun. MORAN: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 90:109–118, 2019. 1, 4, 5, 7

[35] Canjie Luo, Qingxiang Lin, Yuliang Liu, Lianwen Jin, and Chunhua Shen. Separating content from style using adversarial learning for recognizing text in the wild. *Int. J. Comput. Vis.*, 129(4):960–976, 2021. 6

[36] Pengyuan Lyu, Chengquan Zhang, Shanshan Liu, Meina Qiao, Yangliu Xu, Liang Wu, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. MaskOCR: Text recognition with masked encoder-decoder pretraining. *arXiv preprint arXiv:2206.00311*, 2022. 4, 9

[37] Minesh Mathew, Mohit Jain, and CV Jawahar. Benchmarking scene text recognition in devanagari, telugu and malayalam. In *ICDAR*, volume 7, pages 42–46. IEEE, 2017. 3

[38] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *BMVC*. BMVA, 2012. 1, 3

[39] Byeonghu Na, Yoonsik Kim, and Sungrae Park. Multi-modal Text Recognition Networks: Interactive enhancements between visual and semantic features. In *ECCV*, pages 446–463, 2022. 1, 3, 4, 7

[40] Robert Nagy, Anders Dicker, and Klaus Meyer-Wegener. NEOCR: A configurable dataset for natural image text recognition. In *International Workshop on Camera-Based Document Analysis and Recognition*, pages 150–163. Springer, 2011. 3

[41] 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. ICDAR 2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019. In *ICDAR*, pages 1582–1587. IEEE, 2019. 3

[42] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *ICCV*, pages 569–576, 2013. 1, 3

[43] 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. 1, 3, 5

[44] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green AI. *Communications of the ACM*, 63(12):54–63, 2020. 8

[45] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 3, 4

[46] Fenfen Sheng, Zhineng Chen, and Bo Xu. NRTR: A no-recurrence sequence-to-sequence model for scene text recognition. In *ICDAR*, pages 781–786. IEEE, 2019. 1, 4, 7

[47] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2298–2304, 2017. 1, 4, 7

[48] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. ASTER: An attentional scene text recognizer with flexible rectification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(9):2035–2048, 2018. 1, 4, 5, 7

[49] Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. ICDAR 2017 competition on reading chinese text in the wild (rctw-17). In *ICDAR*, volume 1, pages 1429–1434. IEEE, 2017. 3

[50] 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 *CVPR*, pages 8802–8812, 2021. 3

[51] Yipeng Sun, Jiaming Liu, Wei Liu, Junyu Han, Errui Ding, and Jingtuo Liu. Chinese Street View Text: Large-scale chinese text reading with partially supervised learning. In *ICCV*, pages 9086–9095, 2019. 3

[52] 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 *ICDAR*, pages 1557–1562. IEEE, 2019. 3

[53] 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. 3

[54] Zhaoyi Wan, Jielei Zhang, Liang Zhang, Jiebo Luo, and Cong Yao. On vocabulary reliance in scene text recognition. In *CVPR*, pages 11425–11434, 2020. 3, 5

[55] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *ICCV*, pages 1457–1464. IEEE, 2011. 1, 3

[56] Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, and Mingxiang Cai. Decoupled attention network for text recognition. In *AAAI*, volume 34, pages 12216–12224, 2020. 1, 4, 7

[57] Yuxin Wang, Hongtao Xie, Shancheng Fang, Jing Wang, Shenggao Zhu, and Yongdong Zhang. From Two to One: A new scene text recognizer with visual language modeling network. In *CVPR*, pages 14194–14203, 2021. 1, 4, 7

[58] Xudong Xie, Ling Fu, Zhifei Zhang, Zhaowen Wang, and Xiang Bai. Toward Understanding WordArt: Corner-guided transformer for scene text recognition. In *ECCV*, pages 303–321, 2022. 5

[59] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *CVPR*, pages 9653–9663, 2022. 4

[60] Mingkun Yang, Minghui Liao, Pu Lu, Jing Wang, Shenggao Zhu, Hualin Luo, Qi Tian, and Xiang Bai. Reading and Writing: Discriminative and generative modeling for self-supervised text recognition. In *ACM MM*, pages 4214–4223, 2022. 3, 4, 9

[61] Moonbin Yim, Yoonsik Kim, Han-Cheol Cho, and Sungrae Park. Synthtiger: Synthetic text image generator towards better text recognition models. In *ICDAR*, pages 109–124. Springer, 2021. 1

[62] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Towards accurate scene text recognition with semantic reasoning networks. In *CVPR*, pages 12113–12122, 2020. 1, 4, 7

[63] Xiaoyu Yue, Zhanghui Kuang, Chenhao Lin, Hongbin Sun, and Wayne Zhang. RobustScanner: Dynamically enhancing positional clues for robust text recognition. In *ECCV*, pages 135–151. Springer, 2020. 1, 4, 5, 7

[64] Chengwei Zhang, Yunlu Xu, Zhanzhan Cheng, Shiliang Pu, Yi Niu, Fei Wu, and Futai Zou. SPIN: Structure-preserving

inner offset network for scene text recognition. In *AAAI*, volume 35, pages 3305–3314, 2021. 6

[65] 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 *ICDAR*, pages 1577–1581. IEEE, 2019. 3

[66] 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. 3

[67] Yaping Zhang, Shuai Nie, Wenju Liu, Xing Xu, Dongxiang Zhang, and Heng Tao Shen. Sequence-to-sequence domain adaptation network for robust text image recognition. In *CVPR*, pages 2740–2749, 2019. 3

[68] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. EAST: an efficient and accurate scene text detector. In *CVPR*, pages 5551–5560, 2017. 4