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 experiment 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 are 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 SOTA models. 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.

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 cannot 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,
less 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 cannot 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, 58] 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]. Several 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Year</th>
<th>#Original</th>
<th>#Refined</th>
<th>Lang.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEOCR [40]</td>
<td>2011</td>
<td>5K</td>
<td>3K</td>
<td>EN</td>
</tr>
<tr>
<td>Uber-Text [66]</td>
<td>2017</td>
<td>209K</td>
<td>208K</td>
<td>EN</td>
</tr>
<tr>
<td>RCTW [49]</td>
<td>2017</td>
<td>44K</td>
<td>7K</td>
<td>EN, CH</td>
</tr>
<tr>
<td>IIIT-ILST [37]</td>
<td>2017</td>
<td>6K</td>
<td>2K</td>
<td>EN, IN</td>
</tr>
<tr>
<td>MTWI [15]</td>
<td>2018</td>
<td>139K</td>
<td>53K</td>
<td>EN, CN</td>
</tr>
<tr>
<td>COCOTextV2 [53]</td>
<td>2018</td>
<td>201K</td>
<td>73K</td>
<td>EN</td>
</tr>
<tr>
<td>LSVT [51, 52]</td>
<td>2019</td>
<td>382K</td>
<td>38K</td>
<td>EN, CN</td>
</tr>
<tr>
<td>MLT19 [41]</td>
<td>2019</td>
<td>89K</td>
<td>56K</td>
<td>Multi</td>
</tr>
<tr>
<td>ReCTS [65]</td>
<td>2019</td>
<td>109K</td>
<td>25K</td>
<td>EN, CN</td>
</tr>
<tr>
<td>ArT [7]</td>
<td>2019</td>
<td>50K</td>
<td>35K</td>
<td>EN, CN</td>
</tr>
<tr>
<td>IntelOCR [22]</td>
<td>2021</td>
<td>2.57M</td>
<td>2.01M</td>
<td>EN</td>
</tr>
<tr>
<td>TextOCR [50]</td>
<td>2021</td>
<td>822K</td>
<td>568K</td>
<td>EN</td>
</tr>
<tr>
<td>HierText [31]</td>
<td>2022</td>
<td>1.2M</td>
<td>945K</td>
<td>EN</td>
</tr>
</tbody>
</table>

---

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
polygons 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 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Instances</th>
<th># Vocabularies</th>
<th># Vertical Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>MJ+ST</td>
<td>17M</td>
<td>384K</td>
<td>7K</td>
</tr>
<tr>
<td>Union14M-L</td>
<td>4M</td>
<td>707K</td>
<td>110K</td>
</tr>
<tr>
<td>Union14M-U</td>
<td>10M</td>
<td>-</td>
<td>39K</td>
</tr>
</tbody>
</table>

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 large 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.

### 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.

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

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...
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.

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 MAEErec (Sec. 6.4), which can utilize the 10M unlabeled images in Union14M-U through self-supervised pre-training.
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-Union14M 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.
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 ABIEnt[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 de-

![Figure 6. An overview of MAERec. It consists of a ViT as the backbone and an auto-regressive Transformer decoder.](image)

![Figure 7. Reconstruction results on Union14M-L images.](image)

encoder 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.

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.

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-
We report the average accuracy on six common benchmarks.

<table>
<thead>
<tr>
<th>Method</th>
<th>Pre-train</th>
<th>Fine-tune</th>
<th>Acc-CB</th>
<th>Acc-UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>PerSec [28]</td>
<td>100M R</td>
<td>17M S</td>
<td>82.2</td>
<td></td>
</tr>
<tr>
<td>MaskOCR [36]</td>
<td>4.2M R, 100M S</td>
<td>17M S</td>
<td>92.6</td>
<td></td>
</tr>
<tr>
<td>DiG-S [60]</td>
<td>15.8M R, 17M S</td>
<td>2.8M R</td>
<td>94.6</td>
<td></td>
</tr>
<tr>
<td>DiG-B [60]</td>
<td>15.8M R, 17M S</td>
<td>2.8M R</td>
<td>95.0</td>
<td></td>
</tr>
<tr>
<td>MAERec-S (ours)</td>
<td>10.6M R</td>
<td>2.8M R</td>
<td>94.6</td>
<td></td>
</tr>
<tr>
<td>MAERec-S (ours)</td>
<td>10.6M R</td>
<td>3.2M R</td>
<td>96.2</td>
<td></td>
</tr>
<tr>
<td>MAERec-B (ours)</td>
<td>10.6M R</td>
<td>3.2M R</td>
<td>96.2</td>
<td></td>
</tr>
</tbody>
</table>

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.

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).

<table>
<thead>
<tr>
<th>No.</th>
<th>Pre-train</th>
<th>Fine-tune</th>
<th>Acc-CB</th>
<th>Acc-UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>MI, ST</td>
<td>89.9</td>
<td>46.0</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>Union14M-L</td>
<td>94.1</td>
<td>73.5</td>
</tr>
<tr>
<td>3</td>
<td>MI, ST</td>
<td>MI, ST</td>
<td>89.9</td>
<td>46.1</td>
</tr>
<tr>
<td>4</td>
<td>MI, ST</td>
<td>Union14M-L</td>
<td>94.0</td>
<td>75.0</td>
</tr>
<tr>
<td>5</td>
<td>Union14M-U</td>
<td>Union14M-L</td>
<td>95.1</td>
<td>78.6</td>
</tr>
</tbody>
</table>

Table 8. 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. R stands for real data, and S stands for synthetic data. We report the average accuracy on six common benchmarks.

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), Zuhai Industry Core and Key Technology Research Project (no. 2220004002350).

References


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


