Appearance Discrepancy-guided Sequence Hybrid Masking for Robust Scene Text Recognition

Shihao Zou[1,2], **Wei Wei**[1,2*], **Leyang Xu**[2,3], **Kaihe Xu**[2,3], **Wenfeng Xie**[2,3]

[1]School of Computer Science and Technology, Huazhong University of Science and Technology,
[2]Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL),
[3]Ping An Property & Casualty Insurance Company of China, Ltd.
{sh_zou, weiw}@hust.edu.cn, {xuleyang704,xukaihe864,xiewenfeng801}@pingan.com.cn

Abstract

Masked Image Modeling (MIM) has been widely recognized as a powerful self-supervised paradigm for learning general-purpose visual representations. However, standard MIM based on random masking tends to underperform in domain-specific tasks like Scene Text Recognition (STR), due to challenges such as information sparsity and appearance discrepancies caused by partial occlusion or distortion. To address this issue, we propose a novel pre-training framework called **D**iscrepancy-guided **S**equence **H**ybrid **M**asking (DSHM), specifically designed to learn robust representations for STR. To this end, we introduce an Appearance Discrepancy Metric that quantifies the discrepancy level of each image patch by measuring its deviation from anisotropic local discrepancy and intra-instance global style discrepancy. The resulting discrepancy scores are utilized in two key components: (1) A Sequence Hybrid Masking strategy, which prioritizes masking high-discrepancy patches in coherent block forms, thereby elevating the pretext task from simple pixel-level completion to more complex structural reasoning; (2) Discrepancy-Conditioned Tokens (DC-Tokens), which encode prior knowledge about patch difficulty into the decoder, enabling an adaptive reconstruction process and improving the model robustness under scenarios with partial occlusion or text distortion. We achieve competitive performance on multiple benchmark datasets, including common benchmarks, Union14M benchmarks, and Chinese benchmarks.

1 Introduction

Optical character recognition (OCR) is a vital technology for extracting textual information from images, covering tasks such as table structure recognition (Wan et al. 2024), scene text recognition (STR) (Zhao et al. 2024b; Rang et al. 2024), and information extraction from richly formatted document images (Luo et al. 2023; Wang, Xue, and Jin 2024). As a rapidly evolving research area, STR offers numerous opportunities for technological advancements and innovations in fields such as autonomous driving (Zhang et al. 2022), intelligent navigation (Wu et al. 2019), and visual information extraction (Yim et al. 2021).

Currently, prevalent approaches(Jiang et al. 2023; Gao et al. 2024; Zhang et al. 2025) often rely on masked image modeling (MIM) strategies, typically employing random masking of local image regions during pre-training. For example, as illustrated in Fig. 1(a), MAERec(Jiang et al. 2023) utilizes MIM to reconstruct local visual structures, thereby capturing visual features. Although these strategies have demonstrated promising results, their inherent randomness inadequately considers the structural properties and semantic characteristics unique to scene text images, consequently limiting the model's effective exploitation of linguistic knowledge.

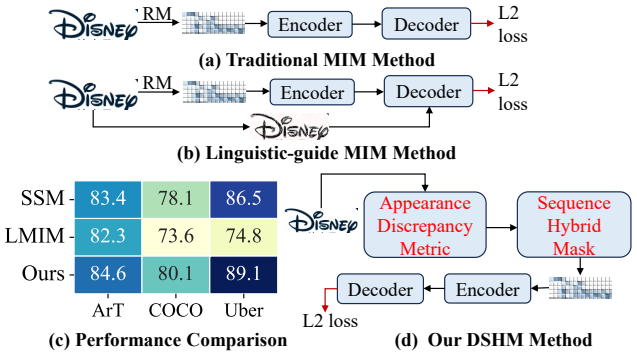

Figure 1: (a),(b),(d) are the comparison with mainstream text recognition methods and our DSHM,(c) is the performance comparison result between ours method and current methods. RM means Random Mask.

Additionally, to more effectively capture local character features and linguistic information within visual contexts, (Gao et al. 2024) proposed the Symmetrical Superimposition Modeling (SSM) approach, which reconstructs pixel-level and feature-level signals from specific directions in input images, thereby learning the visual-linguistic associations within the images. Meanwhile, to better leverage linguistic information during MIM training, LMIM (Zhang et al. 2025) introduced a separate language-guidance branch explicitly integrating linguistic cues into the MIM decoding process. This not only preserves the original visual structure modeling capability inherent in MIM but also significantly enhances the global context-awareness of the model through language guidance, as illustrated in Fig. 1(b). However, the STR domain inherently faces challenges such as informa-

*Corresponding Author.

tion sparsity and appearance discrepancy issues (e.g., curve and occlusion). The random masking strategy used in MIM does not prioritize patches in text images, treating all regions equally. As a result, the model tends to suffer from reduced generalization and robustness when confronted with these challenges. Our comparison with the latest baselines on several complex scenario datasets is illustrated in Fig 1(c).

To address the limitations of current unsupervised methods in STR, we conducted an in-depth analysis of the characteristics of various types of text images. We observed that, compared to standard text, partially occluded or curved text often exhibits appearance discrepancies due to certain characters being obscured or distorted, resulting in inconsistencies with their surrounding context. We refer to this phenomenon as *appearance discrepancy*, which essentially arises from *local feature inconsistency*. Based on this observation, we propose a **D**iscrepancy-guided **S**equence **H**ybrid **M**asking (DSHM) pre-training framework, which guides the model to actively perform discrepancy reasoning, thereby effectively enhancing its robustness and generalization performance under appearance discrepancy challenges. Specifically, we introduce an Appearance Discrepancy Metric (ADM) module to quantify such discrepancies, which measures them from two complementary levels: anisotropic local discrepancy and intra-instance global style discrepancy.

Next, we design a sequence hybrid masking strategy that leverages the ADM-calculated discrepancy scores to prioritize masking of high-scoring, continuous character regions. This guides the model towards learning from more challenging regions. To maintain training stability, a portion of random masking is retained. In the decoding phase, we introduce a Discrepancy-Conditioned Token (DC-Token), which explicitly encodes difficulty signals for the decoder. This enables the decoder to dynamically adapt its reconstruction strategy based on task difficulty, achieving end-to-end discrepancy awareness. Finally, a weighted loss is applied to emphasize learning from these critical regions, compelling the model to invest more capacity in understanding and completing incomplete or irregular visual information.

In summary, the main contributions of this paper are as follows:

- We propose a novel pre-training framework, DSHM, which shifts the pre-training paradigm from passive pattern matching to active, difference reasoning-centered robust learning, offering a unified new perspective for addressing challenges in the STR domain.
- We design an appearance discrepancy metric, which guides the subsequent sequence hybrid masking for more intelligent encoding, and introduces the explicit prior knowledge of DC-Tokens to enable an adaptive reconstruction process.
- The effectiveness of our method is validated on multiple challenging STR benchmark datasets. Experimental results demonstrate that our model achieves superior performance compared to existing state-of-the-art methods.

2 Related Works

Scene Text Recognition

Scene text recognition (STR) methods are typically categorized into language-free and language-aware approaches.

Language-free STR treats recognition as a pure visual task, predicting characters directly from image features without modeling character dependencies. Among them, CTC-based methods (Graves et al. 2006) play a foundational role. Approaches such as CRNN (Shi, Bai, and Yao 2017), GTC (Hu et al. 2020), OrigamiNet (Yousef and Bishop 2020), IFA (Wang et al. 2021a), and PPOCR (Du et al. 2020) use a CTC decoder to convert visual features into text sequences, with blank tokens and post-processing to address alignment. While efficient for inference, their performance is often limited by the lack of contextual modeling.

Language-aware STR incorporates attention mechanisms (Fang et al. 2021; Wang, Da, and Yao 2022; Wang et al. 2021b; Wan et al. 2020) to decode characters by querying visual features at specific positions. These methods are often further enhanced by language modeling, which can be either external (Fang et al. 2021; Yu et al. 2020; Zhao et al. 2024b) or internal (Shi et al. 2019; Cheng et al. 2017; Baek et al. 2019). For example, SRN (Yu et al. 2020) and ABINet (Fang et al. 2021) use external language models to refine predictions, but may miscorrect when visual cues are ignored. In contrast, models like PARSeq (Bautista and Atienza 2022) train internal language models in an autoregressive manner to improve contextual consistency. Recently, BUSNet (Wei et al. 2024) proposed a Balanced, Unified, and Synchronized vision-language framework to reinterpret images as a form of language, mitigating overreliance on either modality.

Self-supervised Learning

In recent years, masked image modeling (MIM) methods, such as MAE (He et al. 2022) and SimMIM (Xie et al. 2022), have attracted increasing attention for their pixel-level reconstruction capabilities. Several works further refine masking strategies: MaskedKD (Son et al. 2024) prioritizes low-attention regions from the student model to emphasize challenging areas during distillation. While these methods effectively leverage sample difficulty, they are task-agnostic and struggle to adapt to domains like scene text, which exhibit strong sequential properties.

To address this, self-supervised learning has been extended to scene text recognition (STR). DIG-SSTR (Yang et al. 2022) combines contrastive and reconstruction-based objectives, inspired by reading and writing processes. CCD (Guan et al. 2023) employs character-level contrastive learning with a spectrum segmentation module. SSM (Gao et al. 2024) reconstructs directional signals from overlapping inputs to enhance structural modeling. LMIM (Zhang et al. 2025) captures intra-character structure and inter-character dependencies with language-guided modeling. Recent works also explore masking improvements tailored to STR: (Tang et al. 2025) adopt multiple language-aware masking schemes (e.g., MLM, PLM) to represent multi-level textual cues; (Yang et al. 2024) unify masked and per-

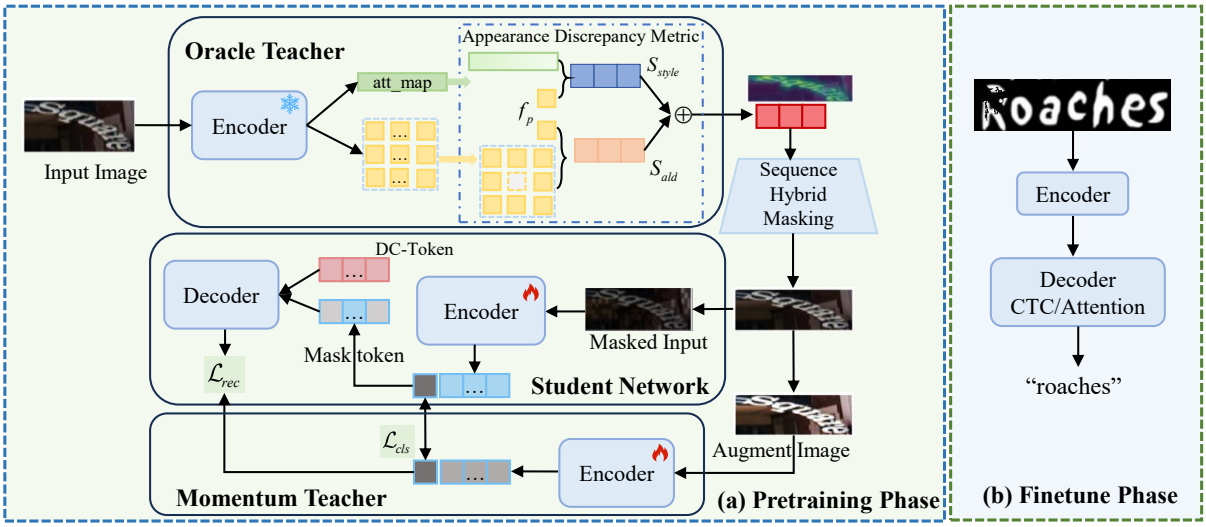

Figure 2: The overall structure of DSHM model.

muted language modeling to enhance contextual reasoning in decoders.

In contrast, we propose a visually grounded pretraining method using an appearance discrepancy metric to identify local inconsistencies and style variations. Unlike general masking, our discrepancy-guided hybrid masking is tailored for sequential text recognition. It encourages the model to attend to structurally or semantically difficult regions during encoding, thus aligning better with the demands of robust STR.

3 Method

Overall Framework

The DSHM framework adopts a carefully designed asymmetric teacher-student architecture, as shown in Figure 2, consisting of a student and two distinct teachers. The Oracle Teacher is a frozen network that provides consistent, high-quality features for computing the appearance discrepancy. The Student Network is the only one updated via gradient descent (parameters θ_s), aiming to reconstruct masked patches using visible context. The Momentum Teacher, updated via exponential moving average of the student (parameters θ_t), offers a stable, evolving reconstruction target. This role separation helps prevent degraded feedback loops during training.

Appearance Discrepancy Metric

Conventional self-supervised methods treat all regions of an image as equally important. However, when applied to the STR task, they overlook a crucial fact: real-world challenges such as occlusion often appear as localized visual discrepancies. We define these explicit differences from standard text images as appearance discrepancies. To equip the model with the ability to recognize and understand such discrepancies, it is essential to first quantify them. To this end, we introduce the Appearance Discrepancy Metric (ADM), a scoring function that quantifies the discrepancy score of each patch p by thoroughly evaluating its relationship with both local structure and global style. The final appearance discrepancy is computed as a weighted combination of these two sub-metrics.

Anisotropic Local Discrepancy: Text is inherently a structural construct, fundamentally determined by the precise arrangement of strokes, junctions, curves, and edges. Simple discrepancy metrics tend to blur these critical details. Our goal is to measure the anisotropy (i.e., direction dependency) of local feature topology.

Specifically, for each patch p, we analyze the set of discrepancy vectors between its feature $f_p \in \mathbb{R}^d$ and those of its eight-connected neighbors $n \in N(p)$. The degree of structural inconsistency can be precisely captured by the total variance within this set of discrepancy vectors, which is formally equivalent to the trace of their covariance matrix. Therefore, the ALD score is defined as:

$$S_{ald}(p) = Tr(Cov(f_p - f_n | n \in N(p))) \quad (1)$$

where Tr denotes the trace operator and Cov represents the covariance. A high S_{ald} score indicates regions with strong feature fluctuations, such as sharp corners, occlusions, or abrupt curves, precisely the areas that demand the most from feature representations.

Intra-Instance Global Style Discrepancy: Discrepancies in text are not limited to geometric structures, they also involve semantic and stylistic variations. A character that deviates in style (e.g., color, font, or brightness) from others within the same word or symbol often represents a critical appearance anomaly that can easily confuse the model. This metric is designed to precisely identify such stylistic or semantic outliers.

Specifically, we leverage mid-level semantic knowledge captured in the Oracle Teacher Attention Map to identify the set of text-related patches within an image instance. These patches exhibit low entropy in their attention distribution, indicating strong structural certainty. We then com-

pute the centroid of these features to form an instance-specific style prototype. Formally, the set is defined as: $\mathcal{T} = \{p|H(A^l(p,\cdot)) < \tau\}$, where A^l denotes the attention map from the l-th layer of the Oracle model, τ is a dynamically determined entropy threshold, and $H(\cdot)$ denotes entropy. We then compute the centroid of these features to form an instance-specific style prototype:

$$\mu_{\mathcal{T}} = \frac{1}{|\mathcal{T}|} \sum_{p \in \mathcal{T}} f_p \tag{2}$$

The style discrepancy score of any patch p is defined as its L2 distance to this contextual prototype, measuring its deviation from the dominant style within the instance:

$$S_{style}(p) = ||f_p - \mu_{\mathcal{T}}||_2 \tag{3}$$

By combining the two metrics, the final ADM score S^{ADM} provides a rich and reliable estimation of patch importance:

$$S^{\text{ADM}}(p) = w_{ald} \cdot S_{ald}(p) + w_{style} \cdot S_{style}(p) \tag{4}$$

where w_{ald} and w_{style} are hyperparameters that balance the contributions of local and global importance.

Sequence Hybrid Masking Strategy

The standard random masking strategy in MIM is "blind" to the specific needs of the STR task, as it disregards the varying semantic importance and recognition difficulty of different image regions. To better align the pre-training objective with the downstream STR task, we argue that masking must be purposeful. It should be: (1) **Appearance Discrepancy-oriented**, focusing on the challenging parts of a text image that most significantly impact model accuracy, and (2) **Task-relevant**, with the masking pattern conforming to the sequential nature of text. Based on these objectives, we have designed a sequence hybrid masking strategy that combines deterministic discrepancy-based masking with random masking.

Specifically, we first rank all patches according to their discrepancy score S^{ADM}, to identify the "most difficult" patches, which then serve as seeds for the subsequent masking process. We generate the difficulty mask by iteratively processing a sorted index list. Furthermore, to simulate the occlusion of continuous characters, we perform block expansion around these "seed" patches. This creates a number of sequential mask blocks, a process that ensures we prioritize masking contiguous blocks of patches that contain high-score seeds.

Finally, to ensure training stability and enable the model to maintain global context awareness, we apply random masking to the remaining unmasked patches. By merging these two sets of masks: discrepancy-based and random, we generate the final mask tensor and the corresponding indices for reconstruction by the decoder. The detailed procedure for implementing this strategy in Algorithm 1.

Algorithm 1: Sequence Hybrid Masking

Require: ADM Scores $S^{\text{ADM}} \in \mathbb{R}^N$ for a sample's N patches; Masking ratios $R_{\text{hard}}, R_{\text{rand}}$; Block size k.
Ensure: Final binary mask $M \in \{0,1\}^N$.

$N_{\text{hard}} \leftarrow \lfloor R_{\text{hard}} \cdot N \rfloor$
$N_{\text{total}} \leftarrow \lfloor (R_{\text{hard}} + R_{\text{rand}}) \cdot N \rfloor$
{*Phase 1: Discrepancy-guided Sequence Masking*}
$I_{\text{seeds}} \leftarrow \text{TopK-Indices}(S^{\text{ADM}}, N_{\text{hard}})$
$I_{\text{hard}} \leftarrow \text{BlockExpand}(I_{\text{seeds}}, k)$
{*Phase 2: Context-aware Random Masking*}
$\mathcal{P}_{\text{avail}} \leftarrow \{1, \ldots, N\} \setminus I_{\text{hard}}$
$N_{\text{rand}} \leftarrow \max(0, N_{\text{total}} - |I_{\text{hard}}|)$
$I_{\text{rand}} \leftarrow \text{RandomSample}(\mathcal{P}_{\text{avail}}, N_{\text{rand}})$
{*Final Combination*}
$I_{\text{final}} \leftarrow I_{\text{hard}} \cup I_{\text{rand}}$
$M \leftarrow \text{IndicesToMask}(I_{\text{final}}, N)$
return M

Discrepancy-Condition Decoder

The standard Masked Autoencoder decoder treats all incoming mask tokens as identical, making it unable to distinguish whether the reconstruction task for a given patch will be simple or difficult. We propose that if the decoder could know the difficulty of a region in advance, it could adopt a more dynamic and intelligent reconstruction strategy.

To achieve this, we introduce a Discrepancy-Conditioned Token (DC-Token). For each masked position i, its input token t_i^{mask}, is no longer a fixed vector. Instead, it is dynamically constructed from a shared base mask embedding, e_{mask}, and an explicit discrepancy embedding$e_{dis,i}$, which is generated from the discrepancy score S^{ADM} associated with the masked patch at position i. The composition is as follows:

$$\begin{aligned} t_j^{\text{mask}} &= e_{\text{mask}} + f_{\theta_d}(S^{\text{ADM}}(j)) \\ \hat{P} &= Decoder(F^{vis}, t_j^{mask}), j \in \mathcal{M} \end{aligned} \tag{5}$$

wheref_{θ_d} is an MLP that maps the scalar discrepancy score into a high-dimensional feature space, F^{vis} are the visible patch features from encoder. This design directly passes the perception learned by the encoder at the front end to the decoder at the back end, thereby enabling end-to-end discrepancy awareness.

Training Objective

The total loss for the pretraining, $\mathcal{L}_{total}$, consists of two components. The primary component is a weighted reconstruction loss $\mathcal{L}_{rec}$, which adjusts the loss weight w_j^{rec} for each masked patch j based on its discrepancy score S_j^{ADM}. This approach ensures that reconstruction errors in more difficult regions are penalized more severely.

$$\mathcal{L}_{rec} = \frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} w_j^{rec} ||\hat{p}_j - t_j||_2^2 \tag{6}$$

where t_j is the reconstruction target, just provide stable, high-quality feature supervision. Additionally, to ensure the

Method	Venue	PT-data	FT-data	Regular Text			Irregular Text			Avg	Params
				IIIT5K	IC13	SVT	IC15	SVTP	CUTE		
				3000	1015	647	1811	645	288		
ABINet (Fang et al. 2021)	CVPR 21	-	U14M-L	97.2	97.2	95.7	87.6	92.1	94.4	94.0	37M
VisionLAN (Wang et al. 2021b)	ICCV 21	-	U14M-L	96.3	95.1	91.3	83.6	85.4	92.4	91.3	33M
SeqCLR (Aberdam et al. 2021)	CVPR 21	U14M-U	U14M-L	98.8	97.9	96.8	91.4	92.9	96.9	95.8	36M
MATRN (Na, Kim, and Park 2022)	ECCV 22	-	U14M-L	98.2	97.9	96.9	88.2	94.1	97.9	95.5	44M
PARSeq (Bautista and Atienza 2022)	ECCV 22	-	U14M-L	98.0	96.8	95.2	85.2	90.5	96.5	93.8	24M
MIM (He et al. 2022)	CVPR 22	U14M-U	U14M-L	98.5	97.6	96.0	89.5	91.3	96.9	95.0	36M
DiG-S (Yang et al. 2022)	ACMMM 22	U14M-U	U14M-L	98.7	97.8	98.5	88.9	92.7	96.5	95.5	36M
CCD (Guan et al. 2023)	ICCV 23	STD&URD	ARD	98.0	98.3	96.4	90.3	92.7	98.3	95.6	36M
MAERec-S (Jiang et al. 2023)	ICCV 23	U14M-U	U14M-L	98.0	97.6	96.8	87.1	93.2	97.9	95.1	36M
MaskOCR-S (Lyu et al. 2024)	TMLR 24	STD&URD	ARD	98.0	97.8	96.9	90.2	94.9	96.2	95.6	31M
SSM-S (Gao et al. 2024)	IJCAI 24	U14M-U	U14M-L	99.0	98.3	97.8	89.5	94.0	98.3	96.1	36M
ViSu* (Qu et al. 2024)	NIPS 24	-	-	98.5	97.8	98.3	90.4	96.3	97.6	96.5	24M
CLIP4STR† (Zhao et al. 2024a)	TIP 24	WIT 400M	Real	99.2	98.3	98.3	91.4	<u>97.2</u>	**99.3**	<u>97.3</u>	158M
BUSNet‡(Wei et al. 2024)	AAAI 24	-	Real	98.0	<u>98.5</u>	<u>98.5</u>	91.3	96.3	98.0	96.8	57M
CSD-S† (Maracani et al. 2025)	CVPR 25	-	Real	<u>99.4</u>	**98.8**	<u>98.5</u>	<u>91.9</u>	**97.5**	<u>99.0</u>	**97.5**	41M
LMIM (Zhang et al. 2025)	CVPR 25	U14M-U	U14M-L	99.3	98.1	98.3	91.7	95.4	**99.3**	97.0	36M
DSHM(Ours)		U14M-U	U14M-L	**99.5**	<u>98.5</u>	**98.7**	**92.3**	96.6	**99.3**	**97.5**	36M

Table 1: Results on six common benchmarks. URD denotes the unlabeled real dataset containing 15.8M images. STD refers to 17M synthetic data. ARD refers to 2.8M annotated real images. Real refers to 3.3M annotated real images. Results marked with $*$ are from (Qu et al. 2024), † are from (Maracani et al. 2025), ‡ are from (Wei et al. 2024), others are from (Zhang et al. 2025). The best results are highlighted in bold, while second-best results are underlined.

model simultaneously learns a high-quality global image representation that is robust to local information loss (i.e., masking), we introduce an auxiliary consistency loss.

$$\mathcal{L}_{cls} = ||v_{stu} - \text{sg}(v_{tea})||_2^2 \tag{7}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. The student vector, $v_{stu} \in \mathbb{R}^d$, is the output from processing the masked image, while the teacher vector comes from processing the original unmasked image and serves as the anchor for global representation alignment. Therefore, the total loss is calculated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda \mathcal{L}_{cls} \tag{8}$$

4 Experiment

Datasets and Evaluation Metrics

Pre-training Data: For English scene text, we use Union14M-U (Jiang et al. 2023), mainly consisting of approximately 10 million unlabeled real-text images extracted and refined from Book32, CC, and OpenImages. Following the LMIM (Zhang et al. 2025) work, we also gratefully use the unlabeled Chinese text image dataset (UCTI-11M) they provided. These datasets are used for model pre-training.

Fine-tuning Data: We fine-tune the model mainly on approximately 3.2 million labeled English text images from the Union14M-L dataset. For Chinese benchmarks (Chen et al. 2021), we use about 1.1 million labeled Chinese images crossing four categories for fine-tuning.

STR Benchmarks: We evaluate our model on commonly used benchmarks including IC13 (Karatzas et al. 2013), SVT (Wang, Babenko, and Belongie 2011), IIIT5K (Mishra, Alahari, and Jawahar 2012), IC15 (Karatzas et al. 2015), SVTP (Phan et al. 2013), and CUTE80 (Risnumawan et al. 2014). Additionally, we assess performance on the real-world Union14M Benchmark. To assess robustness, we conduct experiments on HOST and WOST (Bautista and Atienza 2022) for occluded text. Furthermore, we evaluate the model on three more challenging datasets: COCO (Veit et al. 2016), ArT (Chng et al. 2019), and Uber (Zhang et al. 2017). To test cross-lingual capabilities, we also conduct experiments on a Chinese benchmark consisting of 0.15 million images from four categories (Chen et al. 2021). For English text, we use the standard WAICS metric. For Chinese text, we follow existing evaluation protocols and compute sentence-level accuracy for each subset (Chen et al. 2021).

Performance Comparison

Result on Common Benchmark: Table 1 presents the performance comparison on common benchmarks. The results show that our method achieves a superior average accuracy of 97.5% compared to baseline models trained on similar or comparable amounts of pre-training data. Even when compared to CLIP4STR, which leverages large-scale pre-training data, DSHM remains competitive. Furthermore, our model performs well on irregular datasets, indicating that the proposed discrepancy-guide mechanism enables the model to better understand and reason about challenging regions.

Result on Union14M Benchmark: Considering the performance saturation on common benchmarks, we further conduct evaluations on the more challenging Union14M benchmark, as shown in Table 2. Our method achieves new state-of-the-art accuracy among all recent baselines of similar scale. In particular, DSHM surpasses LMIM by 1.3%

Method	Venue	Datasets	Cur	M-O	Art	Con	Sal	M-W	Gen	AVG	Params
ABINet (Fang et al. 2021)	CVPR 21	-	75.0	61.5	65.3	71.1	72.9	59.1	79.4	69.2	37M
VisionLAN (Wang et al. 2021b)	ICCV 21	-	70.7	57.2	56.7	63.8	67.6	47.3	74.2	62.5	33M
SeqCLR (Aberdam et al. 2021)	CVPR 21	U14M-U+L	83.7	79.9	73.7	79.7	81.0	84.0	82.7	80.7	36M
MATRN (Na, Kim, and Park 2022)	ECCV 22	-	80.5	64.7	71.1	74.8	79.4	67.6	77.9	74.6	44M
PARSeq (Bautista and Atienza 2022)	ECCV 22	-	79.8	79.2	67.4	77.4	77.0	76.9	80.6	76.9	24M
MIM (He et al. 2022)	CVPR 22	U14M-U+L	84.8	79.3	71.1	81.5	81.0	82.5	82.0	80.3	36M
DiG-S (Yang et al. 2022)	ACMMM 22	U14M-U+L	85.9	83.5	77.4	82.5	84.3	84.0	83.8	83.0	36M
MAERec-S (Jiang et al. 2023)	ICCV 23	U14M-U+L	81.4	71.4	72.0	82.0	78.5	82.4	82.5	78.6	36M
SSM-S (Gao et al. 2024)	IJCAI 24	U14M-U+L	87.5	85.8	78.4	84.8	85.2	85.0	84.0	84.3	36M
ViSu† (Qu et al. 2024)	NIPS 24	U14M-U+L	90.7	**96.1**	79.4	85.4	87.1	86.3	83.1	86.9	24M
LMIM (Zhang et al. 2025)	CVPR 25	U14M-U+L	90.3	86.6	80.7	85.5	**88.2**	87.9	85.1	86.3	36M
DSHM(Ours)		U14M-U+L	**91.6**	88.2	**81.4**	**86.3**	87.9	**88.4**	**85.3**	**87.0**	36M

Table 2: Results on English Union14M benchmark. L denotes Union14M-L, multiple baselines were used differently for each of these two datasets. Results marked † are from ViSu (Qu et al. 2024), others are from LMIM (Zhang et al. 2025).

Method	Venue	Pre-train	Scene	Web	Document	Handwriting	Avg	Params
MASTER (Lu et al. 2021)	PR 21	No	62.1	53.4	82.7	18.5	54.2	63M
ABINet (Fang et al. 2021)	CVPR 21	No	60.9	51.1	91.7	13.8	54.4	53M
TransOCR (Chen, Li, and Xue 2021)	CVPR 21	No	67.8	62.7	97.9	51.7	70.0	84M
SVTR-B (Du et al. 2022)	IJCAI 22	No	71.4	64.1	99.3	50.0	71.2	25M
SVTR-L (Du et al. 2022)	IJCAI 22	No	72.1	66.3	99.3	50.3	72.0	41M
CIRN (Yu et al. 2023b)	IJCAI 23	No	73.3	-	-	-	-	-
DCTC-B (Zhang et al. 2024)	AAAI 24	No	72.2	67.0	**99.4**	50.4	72.3	25M
DCTC-L (Zhang et al. 2024)	AAAI 24	No	73.9	68.5	**99.4**	51.0	73.2	41M
TransOCR (Chen, Li, and Xue 2021)	CVPR 21	Yes	68.5	62.5	97.9	53.5	70.6	84M
SeqCLR (Aberdam et al. 2021)	CVPR 21	Yes	81.7	80.5	98.5	60.3	80.3	36M
MIM (He et al. 2022)	CVPR 22	Yes	82.3	80.9	98.9	62.4	81.1	36M
CCR-CLIP (Yu et al. 2023a)	ICCV 23	Yes	71.3	69.2	98.3	60.3	74.8	-
MaskOCR-S (Lyu et al. 2024)	TMLR 24	Yes	71.4	72.5	98.8	55.6	74.6	36M
MaskOCR-B (Lyu et al. 2024)	TMLR 24	Yes	73.9	74.8	99.3	63.7	77.9	100M
LMIM (Zhang et al. 2025)	CVPR 25	Yes	83.6	82.0	99.1	63.9	82.2	36M
DSHM(Ours)		Yes	**85.2**	**82.6**	99.3	**65.0**	**83.0**	36M

Table 3: Results on Chinese benchmark. Baseline results are coming from LMIM(Zhang et al. 2025).

and 0.7% on Curve and Artistic scene, respectively. These improvements validate the effectiveness of DSHM in guiding the model to allocate more attention to difficult text regions during both encoding and decoding, thereby significantly enhancing robustness in complex scenarios.

Result on Chinese Benchmark: Compared to English characters which primarily rely on shape, Chinese characters depend more on stroke structures. To demonstrate the effectiveness of our method in recognizing complex components such as radicals and dense strokes, we evaluate on Chinese benchmarks. As shown in Table 3, our method achieves relatively leading average accuracy, confirming that the model can accurately locate text regions in Chinese images based on inter-patch discrepancy.

Result on Challenge Datasets We evaluate DSHM on five more challenging datasets of varying scales, which primarily consist of weakly/heavily occluded text, irregular text with diverse shapes, and low-resolution scenarios. As shown in Table 4, our method achieves strong performance across these difficult cases, demonstrating that the proposed method is indeed effective in identifying discrepancy or challenging regions and learning to mask them based on appearance discrepancy metrics. This indicates that our model acquires the ability to reason about structural and regional discrepancies already during the pre-training stage, further validating the generalization and robustness of DSHM when dealing with challenging scene text conditions.

Method	HOST 2416	WOST 2416	ArT 34k	COCO 9825	Uber 89.5k
DIG-S(Yang et al. 2022)	72.1	81.1	83.4	78.1	86.5
MAERec*(Jiang et al. 2023)	72.0	83.7	83.1	79.8	85.6
SSM-S(Gao et al. 2024)	–	–	83.4	78.1	86.5
BUSNet(Wei et al. 2024)	–	–	83.4	79.4	83.2
LMIM*(Zhang et al. 2025)	69.6	84.7	83.5	78.5	86.7
DSHM(Ours)	**72.8**	**86.4**	**84.6**	**80.1**	**89.1**

Table 4: Results on more challenge datasets. * means we use publicly released checkpoints to evaluate the method.

Figure 3: Visualization of discrepancy-guided sequence hybrid masking strategy. The first row shows the original input image. The second row displays the heatmap computed from the appearance difference. The third row visualizes all patches identified as high-discrepancy (an intermediate visualization), not the final hybrid mask used in training.

Mask Strategy	HOST	WOST	ArT	COCO	Uber	Avg
SHM	66.1	81.4	83.5	78.2	85.9	**79.0**
HM	65.0	79.2	81.2	76.1	84.4	77.2
RM	62.3	78.3	80.0	75.4	81.3	75.5

Table 5: Comparison of different masking strategies.

	HOST	WOST	ArT	COCO	Uber	Avg
w/o SHM	62.3	78.3	80.0	75.4	81.3	75.5
w/o Weighted-L	65.1	79.6	81.1	77.3	83.3	77.3
w/o C-L	65.6	80.3	82.1	78.3	85.3	78.3
w/o DC-Token	64.1	79.8	80.2	76.5	82.2	76.6

Table 6: Ablation Study. C-L means CLS consistency loss.

Ablation Study

To more effectively evaluate the performance of DSHM, we select two subsets from Union14M-U—Book32 and OpenImages, containing approximately 5 million images.

Sequence Hybrid Masking Strategy. To validate the use of discrepancy-guided SHM, we compare SHM with hybrid masking (HM) without sequence awareness and random masking (RM). As shown in Table 5, SHM performs better than HM by better preserving semantic continuity, as it can mask entire characters in sparse text regions. HM outperforms RM, highlighting the importance of guiding the model to focus on the discrepancies that often affect recognition. Overall, SHM balances character continuity with guidance for difficult regions, leading to superior performance.

Weighted-Loss. As demonstrated in the second row of Table 6, when we replace the weighted loss with a uniform loss, the performance of the model significantly deteriorates. This result clearly indicates that the weighted loss mechanism effectively guides the model to focus explicitly on discrepancy regions caused by image degradation, thereby substantially improving the robustness and generalization ability of the model when facing complex visual scenarios.

CLS Consistency Loss. As shown in the third row of Table 6, removing this loss leads to a slight performance drop. Serving as an effective regularization term, it helps the model learn global semantic representations that are less sensitive to local visual variations, thereby stabilizing and further improving overall performance.

DC-Token. As shown in Table 6, replacing the DC-Token with a static mask token, leads to a notable performance drop. This suggests that the decoder struggles to reconstruct features without being informed of the difficulty perceived by the encoder. Effectively passing this discrepancy-aware signal from the encoder to the decoder completes the feedback loop in our end-to-end framework and is key to enabling adaptive and efficient reconstruction.

H-MR	R-MR	HOST	WOST	ArT	COCO	Uber	Avg
0.6	0.2	64.8	79.9	82.8	76.2	82.6	77.3
0.5	0.3	66.1	81.4	83.5	78.2	85.9	**79.0**
0.4	0.4	63.1	80.0	81.3	77.3	81.5	76.6

Table 7: Comparison of different masking ratios. H, R means Hard and Random, MR means Mask Ratio.

Different Mask Ration Combination. Table 7 shows that the best performance is achieved with HMR = 0.5 and RMR = 0.3, suggesting that a balanced mix of hard and random masking yields the most robust representations. Too much emphasis on either leads to performance degradation.

Effectiveness of Appearance-Guided Masking

We visualize DSHM's pre-training process to evaluate its effectiveness (Fig.3). The second row shows the ADM heatmap. Unlike standard saliency-based attention, our ADM precisely pinpoints difficult regions (e.g., complex strokes or style degradation), validating its unique signal. The third row visualizes all high-discrepancy patches, not the final mask. In practice, we create the hybrid mask by selecting a TopK subset of these patches and combining them with a random mask. This ensures the model is forced to focus on challenging regions while always retaining sufficient low-discrepancy context, promoting structural understanding over trivial pixel inpainting.

5 Conclusion

In this paper, we propose a novel pre-training framework for STR, named DSHM. We first introduce an appearance discrepancy metric to measure the discrepancy level of each image patch from two perspectives: anisotropic local discrepancy and intra-instance global style discrepancy. Based on the resulting scores, we apply a sequence hybrid masking strategy to deterministically mask regions identified as discrepant. Additionally, we design a DC-Token to explicitly convey patch-level difficulty signals to the decoder, enabling a dynamic and adaptive reconstruction process. Guided by appearance discrepancies, our method effectively addresses challenging scenarios such as partial occlusion and curve.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276110, No. 62172039 and in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL). The authors would also like to thank the anonymous reviewers for their comments on improving the quality of this paper.

References

Aberdam, A.; Litman, R.; Tsiper, S.; Anschel, O.; Slossberg, R.; Mazor, S.; Manmatha, R.; and Perona, P. 2021. Sequence-to-Sequence Contrastive Learning for Text Recognition. In *CVPR*, 15302–15312.

Baek, J.; Kim, G.; Lee, J.; Park, S.; Han, D.; Yun, S.; Oh, S. J.; and Lee, H. 2019. What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis. In *ICCV*, 4714–4722.

Bautista, D.; and Atienza, R. 2022. Scene Text Recognition with Permuted Autoregressive Sequence Models. In *ECCV*, volume 13688, 178–196.

Chen, J.; Li, B.; and Xue, X. 2021. Scene Text Telescope: Text-Focused Scene Image Super-Resolution. In *CVPR*, 12026–12035.

Chen, J.; Yu, H.; Ma, J.; Guan, M.; Xu, X.; Wang, X.; Qu, S.; Li, B.; and Xue, X. 2021. Benchmarking Chinese Text Recognition: Datasets, Baselines, and an Empirical Study. *arXiv*, abs/2112.15093.

Cheng, Z.; Bai, F.; Xu, Y.; Zheng, G.; Pu, S.; and Zhou, S. 2017. Focusing Attention: Towards Accurate Text Recognition in Natural Images. In *ICCV*, 5086–5094.

Chng, C. K.; Liu, Y.; Sun, Y.; Ng, C. C.; Luo, C.; Ni, Z.; Fang, C.; Zhang, S.; Han, J.; Ding, E.; et al. 2019. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *ICDAR*, 1571–1576.

Du, Y.; Chen, Z.; Jia, C.; Yin, X.; Zheng, T.; Li, C.; Du, Y.; and Jiang, Y. 2022. SVTR: Scene Text Recognition with a Single Visual Model. In *IJCAI*, 884–890.

Du, Y.; Li, C.; Guo, R.; Yin, X.; Liu, W.; Zhou, J.; Bai, Y.; Yu, Z.; Yang, Y.; Dang, Q.; et al. 2020. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*.

Fang, S.; Xie, H.; Wang, Y.; Mao, Z.; and Zhang, Y. 2021. Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition. In *CVPR*, 7098–7107.

Gao, Z.; Wang, Y.; Qu, Y.; Zhang, B.; Wang, Z.; Xu, J.; and Xie, H. 2024. Self-Supervised Pre-training with Symmetric Superimposition Modeling for Scene Text Recognition. In *IJCAI*, 767–775.

Graves, A.; Fernández, S.; Gomez, F. J.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 369–376.

Guan, T.; Shen, W.; Yang, X.; Feng, Q.; Jiang, Z.; and Yang, X. 2023. Self-Supervised Character-to-Character Distillation for Text Recognition. In *ICCV*, 19473–19484.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. B. 2022. Masked Autoencoders Are Scalable Vision Learners. In *CVPR*, 15979–15988. IEEE.

Hu, W.; Cai, X.; Hou, J.; Yi, S.; and Lin, Z. 2020. GTC: Guided Training of CTC towards Efficient and Accurate Scene Text Recognition. In *AAAI*, 11005–11012.

Jiang, Q.; Wang, J.; Peng, D.; Liu, C.; and Jin, L. 2023. Revisiting Scene Text Recognition: A Data Perspective. In *ICCV*, 20486–20497.

Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V. R.; Lu, S.; et al. 2015. ICDAR 2015 competition on robust reading. In *ICDAR*, 1156–1160.

Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; i Bigorda, L. G.; Mestre, S. R.; Mas, J.; Mota, D. F.; Almazan, J. A.; and De Las Heras, L. P. 2013. ICDAR 2013 robust reading competition. In *ICDAR*, 1484–1493.

Lu, N.; Yu, W.; Qi, X.; Chen, Y.; Gong, P.; Xiao, R.; and Bai, X. 2021. MASTER: Multi-aspect non-local network for scene text recognition. *PR*, 117: 107980.

Luo, C.; Cheng, C.; Zheng, Q.; and Yao, C. 2023. GeoLayoutLM: Geometric Pre-training for Visual Information Extraction. In *CVPR*, 7092–7101.

Lyu, P.; Zhang, C.; Liu, S.; Qiao, M.; Xu, Y.; Wu, L.; Yao, K.; Han, J.; Ding, E.; and Wang, J. 2024. MaskOCR: Scene Text Recognition with Masked Vision-Language Pre-training. *TMLR*.

Maracani, A.; Özkan, S.; Cho, S.; Kim, H.; Noh, E.; Min, J.; Min, C. J.; Park, D.; and Ozay, M. 2025. Accurate Scene Text Recognition with Efficient Model Scaling and Cloze Self-Distillation. *CoRR*, abs/2503.16184.

Mishra, A.; Alahari, K.; and Jawahar, C. 2012. Scene text recognition using higher order language priors. In *BMVC-British machine vision conference*. BMVA.

Na, B.; Kim, Y.; and Park, S. 2022. Multi-modal Text Recognition Networks: Interactive Enhancements Between Visual and Semantic Features. In *ECCV*, volume 13688, 446–463. Springer.

Phan, T. Q.; Shivakumara, P.; Tian, S.; and Tan, C. L. 2013. Recognizing text with perspective distortion in natural scenes. In *ICCV*, 569–576.

Qu, Y.; Wang, Y.; Zhou, B.; Wang, Z.; Xie, H.; and Zhang, Y. 2024. Boosting Semi-Supervised Scene Text Recognition via Viewing and Summarizing. In *NIPS*.

Rang, M.; Bi, Z.; Liu, C.; Wang, Y.; and Han, K. 2024. An Empirical Study of Scaling Law for Scene Text Recognition. In *CVPR*, 15619–15629.

Risnumawan, A.; Shivakumara, P.; Chan, C. S.; and Tan, C. L. 2014. A robust arbitrary text detection system for natural scene images. *ESWA*, 41(18): 8027–8048.

Shi, B.; Bai, X.; and Yao, C. 2017. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. *TPAMI*, 39(11): 2298–2304.

Shi, B.; Yang, M.; Wang, X.; Lyu, P.; Yao, C.; and Bai, X. 2019. ASTER: An Attentional Scene Text Recognizer with Flexible Rectification. *TPAMI*, 41(9): 2035–2048.

Son, S.; Ryu, J.; Lee, N.; and Lee, J. 2024. The Role of Masking for Efficient Supervised Knowledge Distillation of Vision Transformers. In *ECCV*, volume 15125, 379–396. Springer.

Tang, Z.; Mitsui, Y.; Miyazaki, T.; and Omachi, S. 2025. Joint Low-level and High-level Textual Representation Learning with Multiple Masking Strategies. *CoRR*, abs/2505.06855.

Veit, A.; Matera, T.; Neumann, L.; Matas, J.; and Belongie, S. 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*.

Wan, J.; Song, S.; Yu, W.; Liu, Y.; Cheng, W.; Huang, F.; Bai, X.; Yao, C.; and Yang, Z. 2024. OMNIPARSER: A Unified Framework for Text Spotting, Key Information Extraction and Table Recognition. In *CVPR*, 15641–15653.

Wan, Z.; He, M.; Chen, H.; Bai, X.; and Yao, C. 2020. TextScanner: Reading Characters in Order for Robust Scene Text Recognition. In *AAAI*, 12120–12127.

Wang, K.; Babenko, B.; and Belongie, S. 2011. End-to-end scene text recognition. In *ICCV*, 1457–1464.

Wang, P.; Da, C.; and Yao, C. 2022. Multi-granularity Prediction for Scene Text Recognition. In *ECCV*, volume 13688, 339–355. Springer.

Wang, R.; Xue, Y.; and Jin, L. 2024. DocNLC: A Document Image Enhancement Framework with Normalized and Latent Contrastive Representation for Multiple Degradations. In *AAAI*, 5563–5571.

Wang, T.; Zhu, Y.; Jin, L.; Peng, D.; Li, Z.; He, M.; Wang, Y.; and Luo, C. 2021a. Implicit Feature Alignment: Learn To Convert Text Recognizer to Text Spotter. In *CVPR*, 5973–5982.

Wang, Y.; Xie, H.; Fang, S.; Wang, J.; Zhu, S.; and Zhang, Y. 2021b. From Two to One: A New Scene Text Recognizer with Visual Language Modeling Network. In *ICCV*, 14174–14183.

Wei, J.; Zhan, H.; Lu, Y.; Tu, X.; Yin, B.; Liu, C.; and Pal, U. 2024. Image as a Language: Revisiting Scene Text Recognition via Balanced, Unified and Synchronized Vision-Language Reasoning Network. In *AAAI*, 5885–5893. AAAI Press.

Wu, L.; Zhang, C.; Liu, J.; Han, J.; Liu, J.; Ding, E.; and Bai, X. 2019. Editing Text in the Wild. In *ACM MM*, 1500–1508.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. SimMIM: a Simple Framework for Masked Image Modeling. In *CVPR*, 9643–9653. IEEE.

Yang, M.; Liao, M.; Lu, P.; Wang, J.; Zhu, S.; Luo, H.; Tian, Q.; and Bai, X. 2022. Reading and Writing: Discriminative and Generative Modeling for Self-Supervised Text Recognition. In *ACM MM*, 4214–4223.

Yang, X.; Qiao, Z.; Wei, J.; Yang, D.; and Zhou, Y. 2024. Masked and Permuted Implicit Context Learning for Scene Text Recognition. *IEEE Signal Process. Lett.*, 31: 964–968.

Yim, M.; Kim, Y.; Cho, H.-C.; and Park, S. 2021. Synthtiger: Synthetic text image generator towards better text recognition models. In *ICDAR*, 109–124.

Yousef, M.; and Bishop, T. E. 2020. OrigamiNet: Weakly-Supervised, Segmentation-Free, One-Step, Full Page Text Recognition by learning to unfold. In *CVPR*, 14698–14707.

Yu, D.; Li, X.; Zhang, C.; Liu, T.; Han, J.; Liu, J.; and Ding, E. 2020. Towards Accurate Scene Text Recognition With Semantic Reasoning Networks. In *CVPR*, 12110–12119.

Yu, H.; Wang, X.; Li, B.; and Xue, X. 2023a. Chinese Text Recognition with A Pre-Trained CLIP-Like Model Through Image-IDS Aligning. In *ICCV*, 11909–11918.

Yu, H.; Wang, X.; Li, B.; and Xue, X. 2023b. Orientation-Independent Chinese Text Recognition in Scene Images. In *IJCAI*, 1667–1675.

Zhang, C.; Tao, Y.; Du, K.; Ding, W.; Wang, B.; Liu, J.; and Wang, W. 2022. Character-Level Street View Text Spotting Based on Deep Multisegmentation Network for Smarter Autonomous Driving. *TAI*, 3(2): 297–308.

Zhang, Y.; Gueguen, L.; Zharkov, I.; Zhang, P.; Seifert, K.; and Kadlec, B. 2017. Uber-text: A large-scale dataset for optical character recognition from street-level imagery. In *CVPRW*, volume 2017, 5.

Zhang, Y.; Liu, C.; Wei, J.; Yang, X.; Zhou, Y.; Ma, C.; and Ji, X. 2025. Linguistics-aware Masked Image Modeling for Self-supervised Scene Text Recognition. *CoRR*, abs/2503.18746.

Zhang, Z.; Lu, N.; Liao, M.; Huang, Y.; Li, C.; Wang, M.; and Peng, W. 2024. Self-Distillation Regularized Connectionist Temporal Classification Loss for Text Recognition: A Simple Yet Effective Approach. In *AAAI*, 7441–7449.

Zhao, S.; Quan, R.; Zhu, L.; and Yang, Y. 2024a. CLIP4STR: A Simple Baseline for Scene Text Recognition With Pre-Trained Vision-Language Model. *TIP*, 33: 6893–6904.

Zhao, Z.; Tang, J.; Lin, C.; Wu, B.; Huang, C.; Liu, H.; Tan, X.; Zhang, Z.; and Xie, Y. 2024b. Multi-modal In-Context Learning Makes an Ego-evolving Scene Text Recognizer. In *CVPR*, 15567–15576.