The Devil is in Fine-tuning and Long-tailed Problems: A New Benchmark for Scene Text Detection

Tianjiao Cao[1,3], **Jiahao Lyu**[1,3], **Weichao Zeng**[1,3], **Weimin Mu**[1] and **Yu Zhou**[2,✉]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]VCIP & TMCC & DISSec, College of Computer Science, Nankai University
[3]School of Cyber Security, University of Chinese Academy of Sciences
{caotianjiao, lvjiahao, zengweichao, muweimin}@iie.ac.cn, yzhou@nankai.edu.cn

Abstract

Scene text detection has seen the emergence of high-performing methods that excel on academic benchmarks. However, these detectors often fail to replicate such success in real-world scenarios. We uncover two key factors contributing to this discrepancy through extensive experiments. First, a *Fine-tuning Gap*, where models leverage *Dataset-Specific Optimization* (DSO) paradigm for one domain at the cost of reduced effectiveness in others, leads to inflated performances on academic benchmarks. Second, the suboptimal performance in practical settings is primarily attributed to the long-tailed distribution of texts, where detectors struggle with rare and complex categories as artistic or overlapped text. Given that the DSO paradigm might undermine the generalization ability of models, we advocate for a *Joint-Dataset Learning* (JDL) protocol to alleviate the Fine-tuning Gap. Additionally, an error analysis is conducted to identify three major categories and 13 subcategories of challenges in long-tailed scene text, upon which we propose a Long-Tailed Benchmark (LTB). LTB facilitates a comprehensive evaluation of ability to handle a diverse range of long-tailed challenges. We further introduce MAEDet, a self-supervised learning-based method, as a strong baseline for LTB. The code is available at https://github.com/pd162/LTB.

1 Introduction

Scene Text Detection (STD) has received extensive attention both in academia and industry due to its wide applications, such as scene understanding [Zhang *et al.*, 2025; Lyu *et al.*, 2025], intelligent office [Zeng *et al.*, 2023; Shen *et al.*, 2024; Shen *et al.*, 2023], etc. Researchers have proposed many prominent scene text detectors [Zhou *et al.*, 2017; Long *et al.*, 2018; Wang *et al.*, 2019; Liao *et al.*, 2020; Liu *et al.*, 2021; Du *et al.*, 2022; Qin *et al.*, 2023; Bu *et al.*, 2024]. Figure 1 (top) shows the year-over-year performance improvement of scene text detectors across three public benchmarks. Recent advances in scene text detection show consistent improvements in F-measure scores, with recent models surpassing 90%, reflecting rapid progress in the field.

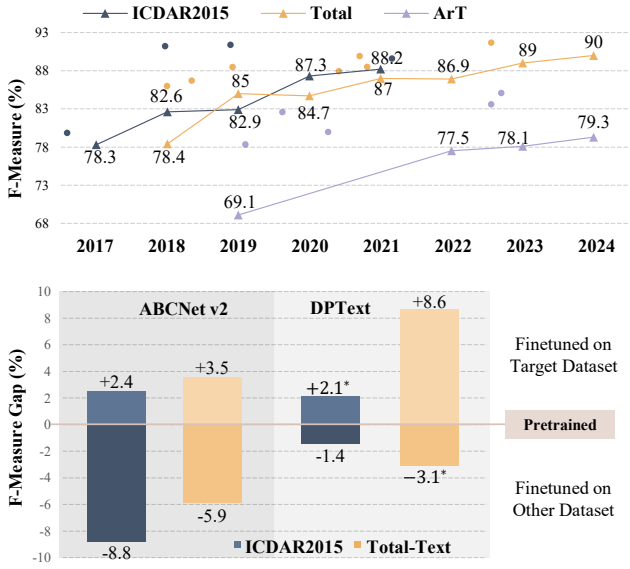

Figure 1: (top) The performance tendency of scene text detectors in recent years. We choose several representative methods in chronological order. Circles indicate the performances of other methods. (bottom) A visualization of fine-tuning gap. Each bar shows a performance comparison on the same benchmark. "$*$" denotes we fine-tuned the model with the open-source pretrained weight.

Despite the impressive benchmark results, scene text detectors still face challenges (e.g., texts within bad illumination and complex backgrounds) and fail to meet the expectations in real-world applications. The significant discrepancy between the high performance on academic benchmarks and the subpar performance in practical deployments prompts our investigation into the underlying causes. We attribute this to two main factors: fine-tuning gap and long-tailed problem.

Firstly, previous evaluations of scene text detectors mainly rely on the *Dataset-Specific Optimization* (DSO) paradigm, where detectors are pretrained on a variety of datasets and then fine-tuned on a single dataset's training subset, followed by assessment on its corresponding test subset. However, such a protocol might not reflect authentic generalization

✉ Corresponding Author.

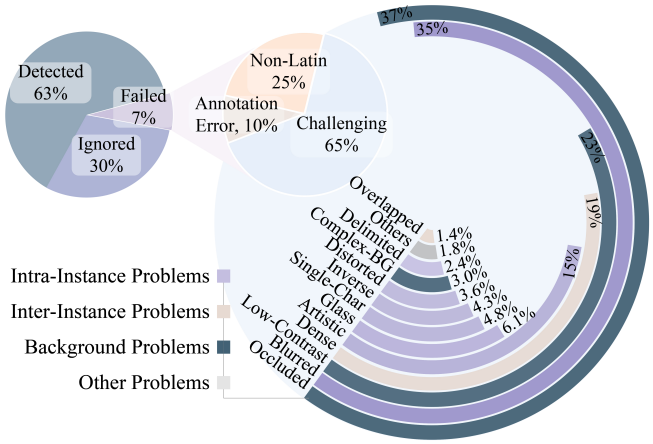

Figure 2: A quantitative analysis of failure cases. The derivation of results is described in Section 4. The left figure shows that the overall detection results are divided into ignored, detected, and failed types. The middle presents statistics of failure cases with human annotation. The right illustrates the distribution of challenging text instances. The brightness of color indicates the prevalence of each category, with darker colors representing higher proportions. BG for background and Char for character.

ability, as each dataset exhibits unique biases and domain-specific characteristics. Fine-tuning on a biased training subset could lead to overfitting to a specific domain, resulting in degraded performance on others. As shown in Figure 1 (bottom), a model consistently improves its performance on the test set after fine-tuning with the corresponding training data. Conversely, its performance drops significantly, with a decrease of up to 8.8% when fine-tuned on one dataset and evaluated on another. Evaluations on limited benchmarks reflect the detector's effectiveness for specific text types but fail to assess its overall performance in diverse scenarios. Previous works [Zhan *et al.*, 2019; Tian *et al.*, 2022] also refer to the domain-adaptation problem in STD. To address this issue, we propose a *Joint-Dataset Learning* (JDL) method, where a detector is trained on a unified training set from multiple datasets and directly evaluated on a combined test set. We posit that JDL offers a more accurate and comprehensive performance assessment.

Secondly, scene text categories commonly exhibit a long-tailed distribution in real-world scenarios, with dominant head categories (e.g., printed text) and scarce tail categories (e.g., artistic text). We define long-tailed problems as challenges where detectors struggle with tail categories due to the lack of relevant training samples, resulting in substantial performance degradation on less frequent but critical text instances in practical applications. However, most existing studies focus on specific challenges and overlook long-tailed issues from a holistic perspective. In light of this, we analyze and define a comprehensive set of challenges, as shown in Figure 2. We evaluate four representative scene text detectors (CNN-based and Transformer-based) on three datasets with distinct styles (for diverse scenarios). Results reveal that the detectors fail to identify 7% of text instances. Among these undetected instances, only 65% are found to be genuinely challenging, while the remaining 35% are attributed to annotation errors or involved non-Latin scripts. We identify recurring patterns across these challenging cases and categorize them into three main groups: intra-instance, inter-instance, and background problems. These groups are further divided into 13 subcategories, each representing a long-tailed challenge in scene text detection. To support future research, we introduce a Long-Tailed Benchmark (LTB) as a standardized platform for evaluating the ability of scene text detectors to address these challenges.

Finally, inspired by MAERec [Jiang *et al.*, 2023], we utilize self-supervised learning (SSL) for better text representation and propose a baseline based on MAE [He *et al.*, 2022] from the data perspective, called MAEDet. In conclusion, our contributions are three-fold.

- We identify the devil of the discrepancy between the performance on academic benchmarks and that on real applications lies in fine-tuning and long-tailed problems.
- We introduce the *Joint-Dataset Learning* (JDL) protocol as a replacement for *Dataset-Specific Optimization* to avoids the fine-tuning gap and provides a more accurate evaluation of model generalization across different domains.
- We propose the Long-Tailed Benchmark (LTB), which poses 13 distinct challenges and serves as the first benchmark to comprehensively evaluate models' capability to address various long-tailed problems in scene text detection. A baseline based on self-supervised learning for the benchmark is also provided.

2 Related Works

2.1 Approaches for Scene Text Detection

CNN-based methods. [Gupta *et al.*, 2016; Jaderberg *et al.*, 2016; Liao *et al.*, 2017; Liao *et al.*, 2018] are followed the object detection method to detect the scene text. [He *et al.*, 2017] and its variants [Xiao *et al.*, 2020; Du *et al.*, 2022; Ye *et al.*, 2020] represent the oriented and curve text contours. [Long *et al.*, 2018; Liu *et al.*, 2020; Wang *et al.*, 2020; Wang *et al.*, 2022; Liu *et al.*, 2023; Su *et al.*, 2024] also explore the efficient representation operators.

Transformer-based methods. Attributing to the broad application in visual tasks, Transformer, as the fundamental architecture, has been gradually embedded in the scene text detector [Raisi *et al.*, 2021; Tang *et al.*, 2022; Liu *et al.*, 2023; Bu *et al.*, 2024]. [Raisi *et al.*, 2021] firstly introduces the Deformable DETR [Zhu *et al.*, 2021] to detect oriented text instances. To alleviate the burden of the Transformer, [Tang *et al.*, 2022] samples the features of text instances and reduces the computational cost. [Ye *et al.*, 2023] notices the reading of inverse-like scene text. [Bu *et al.*, 2024] combines the advantages of top-down and bottom-up and proposes a transformer-powered text detector.

2.2 Benchmarks for Scene Text Detection

Thanks to the Robust Reading Contest, massive and diversified datasets are proposed and evaluated fairly, such as ICDAR 2013 [Karatzas *et al.*, 2013] for axis-align texts, ICDAR

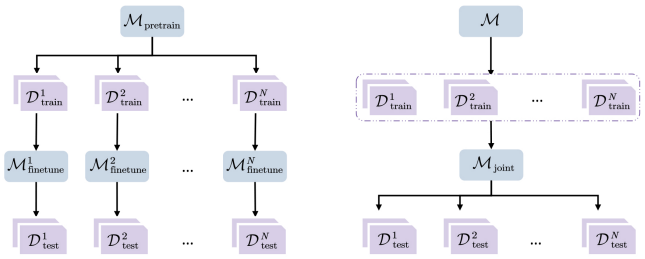

Figure 3: Comparison of the Dataset-Specific Optimization (left) and Joint-Dataset Learning (right) paradigms.

Method	IC15		TT		IC15	TT →
	P	F	P	F	→ TT	IC15
ABCNet v2	86.2†	88.2†	83.7†	87.2†	77.8	77.4
DPText-DETR	75.3	77.4	80.4	89.0†	77.3	73.9

Table 1: Statistics for fine-tuning gap. F-measure (%) is reported with "P" for the pretrained model and "F" for the fine-tuned model. "A → B" denotes fine-tuning on the training set of A and evaluation on test set of B. "†" denotes the official results from the corresponding paper.

2015 [Karatzas *et al.*, 2015] for oriented and unfocused texts, ArT [Chng *et al.*, 2019] for arbitrarily shaped texts, and MLT [Nayef *et al.*, 2017; Nayef *et al.*, 2019] for multi-language texts. These datasets include thousands of text images for training and testing. LSVT [Sun *et al.*, 2019] and TextOCR [Singh *et al.*, 2021] collect larger-scale annotated datasets. Additionally, some long-tailed problems in scene text detection have been noticed in recent years. For example, SCUT-CTW1500 [Liu *et al.*, 2017] is designed to handle curved text with line-level granularity. DAST1500 [Tang *et al.*, 2019], HierText [Long *et al.*, 2022], and InverseText [Ye *et al.*, 2023] solve the problems in dense, multi-hierarchical and inversed texts, respectively. Our current research focuses on English word-level texts, with prospects for future expansion to additional languages and varied granularity.

3 Addressing the Fine-Tuning Gap

Traditional scene text detection methods often adopt the DSO paradigm, which can be formulated as follows. Given N datasets $\mathcal{D} = \{\mathcal{D}^i \mid i = 1, 2, \cdots, N\}$, where each dataset $\mathcal{D}^i$ consists of a training set $\mathcal{D}^i_{\text{train}}$ and a testing set $\mathcal{D}^i_{\text{test}}$, a pretrained model $\mathcal{M}_{\text{pretrain}}$ is first obtained by training on large-scale data (typically synthetic data or a combination of multiple real datasets). The pretrained model is then fine-tuned separately on each training set $\mathcal{D}^i_{\text{train}}$ to produce N fine-tuned models $\mathcal{M}^i_{\text{finetune}}$, which are finally evaluated on their corresponding test sets $\mathcal{D}^i_{\text{test}}$, as illustrated in Figure 3 (left).

However, different benchmarks are tailored to solve various issues (e.g., oriented text detection, curved text detection) and exhibit significant variability due to diverse scenarios and data collection methods, leading to substantial dataset biases. A comprehensive analysis of these variations is provided in Appendix. This issue is further exacerbated by the DSO protocol, where fine-tuning a model on a single training set $\mathcal{D}^i_{\text{train}}$ might not adequately capture the overall distribution of real-world scenarios, leading to performance degradation on other test sets $\mathcal{D}^{\setminus i}_{\text{test}}$.

We define this limitation as the *Fine-tuning Gap*: the discrepancy between a model's performance on its fine-tuned dataset and its ability to generalize to unseen data distributions. To verify the fine-tuning gap, we conduct experiments using CNN-based ABCNet v2 and transformer-based DPText-DETR, with the results presented in Table 1. As shown in the first two columns, fine-tuning yields at least a 2.0% improvement on the target benchmark. However, the last two columns reveal degraded performance when models are fine-tuned on another dataset, which occurs irrespective of the detector architecture, whether CNN-based or Transformer-based. These findings underscore that the DSO approach is ill-suited for actual applications due to several critical issues: 1) Lack of generalization. Detectors fine-tuned on a single dataset struggle to adapt to real-world scenarios' diverse and variable conditions, often resulting in overfitting. 2) Resource inefficiency. Separate fine-tuning for each dataset requires substantial computational and time resources, significantly limiting scalability and practicality. 3) Unrepresentative benchmarking: Evaluating a model on a single, limited test set fails to provide an objective and comprehensive assessment, potentially leading to misleading conclusions.

To address the aforementioned issues, we introduce *Joint-Dataset Learning* (JDL) protocol to STD, as illustrated in Figure 3 (right). JDL draws inspiration from previous advancements in scene text recognition and offers a feasible solution from data perspective. The protocol involves training a model $\mathcal{M}_{\text{joint}}$ on the combined diversity of multiple datasets $\bigcup_{i=1}^{N} \mathcal{D}^i_{\text{train}}$, and directly evaluating it on $\bigcup_{i=1}^{N} \mathcal{D}^i_{\text{test}}$. This unified paradigm offers a more efficient training framework and a more equitable assessment mechanism for STD in practical applications.

We suggest that this new protocol provides a more comprehensive and accurate reflection of a model's adaptability to the complexities of real-world scenarios. To provide a more comprehensive view of the advantages and implications of Joint-Dataset Learning, we delve deeper into its importance and feasibility in Appendix.

4 A Long-tailed Benchmark

Except for the fine-tuning gap, scene text detection also suffers from the problem of long-tailed distribution. Text instances of tailed categories rarely exist in the training set, bringing obstacles to vanilla text detectors. In this section, we will define the common problems of current detection systems, and introduce our approach to design a benchmark to address them.

4.1 Challenge Definition

To uncover the challenges encountered in scene text detection, we first apply the four representative text detectors mentioned in Table 2 to the test sets of ICDAR2015 [Karatzas

https://github.com/Yuliang-Liu/bezier_curve_text_spotting
https://github.com/ymy-k/DPText-DETR

Figure 4: An overview of LTB which collects the challenging issues in scene text detection, involving the *Intra-Instance Problems*, *Inter-Instance Problems*, and *Background Problems*. Except for the *others* category, there are 12 kinds of challenges in total, each is illustrated with two images. Left: a complete scene featuring challenging text instances, highlighted by red masks and arrows pointing to their locations. Right: a series of cropped challenging text instances, with the corresponding instance from the left image highlighted by a bold border.

et al., 2015], Total-Text [Ch'ng and Chan, 2017], and ArT [Chng *et al.*, 2019]. By analyzing the failure cases of these detectors, we identify recurring features among the challenging texts. Based on the comprehensive analysis, we categorize the undetected issues into three main groups: intra-instance problems, inter-instance problems, and background problems. Each category poses specific issues that impact text detection in distinct ways. Figure 4 visualizes these categories and their relationships.

Intra-instance Problems

Intra-instance problems refer to challenges associated with the internal attributes of independent text instances. In other words, these problems derive from the low-quality appearance of texts rather than influence from adjacent texts or surrounding backgrounds.

Blurred Text. Blurred text may arise from out-of-focus conditions, like camera shake, limited camera performance, or image noise. Especially in scene images, texts are often not dominant and easy to be unfocused. As shown in Figure 4(a), blurred text is prevalent in scene text.

Artistic Text. Artistic texts frequently appear in graffiti, posters, and advertisements, characterized by elaborate but unconventional designs. These texts may feature highly unique fonts, diverse colors, and non-traditional character arrangements, such as staggered placements where characters are not aligned in a straight line. The non-standard instinct of artistic texts makes it difficult for detectors to generalize from training data, necessitating the development of advanced algorithms to handle these diverse forms effectively.

Glass Text. Glass texts refer to texts that are either directly on the surface of glass or reflected onto it. The optical properties of glass, including reflection, refraction, and distortion, present unique challenges for text detection.

Single-Character Text. In practical applications, individual characters may be large and clear yet still hard to detect because the lack of surrounding contextual clues from adjacent characters or words prevents the detection process.

Distorted Text. Text distortion may arise from several factors: capturing text from highly oblique angles, applying perspective transformations to the text image, or the existence of texts on irregular surfaces. Such misalignment and warping alter the original text proportions and geometric features, creating substantial challenges for detection systems.

Inverse Text. Inverse text occurs when a surface is visible from both sides, with one side showing the text normally and the other side displaying it in reverse. It can also result from image processing techniques like horizontal flipping. DPText-DETR[Ye *et al.*, 2023] argues that inverse-like text instances harm the robustness of the scene text detector.

Delimited Text. A delimited text instance contains formats where specific delimiters segment the content, resulting in multiple discrete pieces rather than a unified whole. Examples include URLs (e.g., www.example.com), phone numbers (e.g., tel:123-4567), and numerical sequences (e.g., 1,0978). The segmentation can result in detection systems losing smaller delimiters and treating the segments as separate text instances, complicating the detection and analysis of overall content.

Inter-instance Problems

In addition to the challenges posed by individual text instances, the interaction among adjacent texts also exerts an influence. Specifically, dense and overlapped texts complicate text detection processes.

Dense Text. Dense text refers to instances where text is closely positioned, either in multi-line or side-by-side formats. This closeness may result in the incorrect merging of

	Dataset	DBNet++	ABCNet v2	DPText-DETR	SRFormer
Detector	ICDAR2015	✓	✓		
	Total-Text	✓	✓	✓	✓
	ArT	✓		✓	✓
Manual	Inverse-Text	-	-	-	-
	NightTime-ArT	-	-	-	-
	ORT	-	-	-	-

Table 2: Dual strategy of dataset collection. "✓" indicates detector is used for processing the dataset. "ORT" refer to Occluded RoadText.

separate text instances into one, which is a common issue in document-like images.

Overlapped Text. When text instances overlap, they become partially obstructed but remain visible. This commonly occurs due to shadows cast on the text, reflections from glossy surfaces, or the presence of watermarks.

Background Problems

The challenges associated with noisy backgrounds stem from intricate or distracting patterns and colors. When backgrounds compete with text for visual attention, it complicates the detection process.

Occluded Text. Occluded text occurs when text instances are partially invisible due to being obscured by foreground objects, excessive illumination, or being located at the edge of an image where parts of the text extend beyond the image boundary. This phenomenon commonly arises in natural scenes, resulting in missing segments.

Low-Contrast Text. Low-contrast text refers to text that has a low contrast with its background. This typically occurs when the text color is similar to the background color or when the text is inadequately illuminated. Such text is challenging for humans to detect even in high-quality images, making it even more difficult for detection algorithms to identify.

Complex-Background Text.

It is difficult to discern the text boundaries when the background contains patterns that closely resemble the text or when the background features complex, multi-colored designs. For instance, text overlaid on highly textured or patterned backdrops or hollowed text that is outlined rather than filled tends to confuse text detectors.

Other Problems

Additionally, some challenges remain difficult to categorize because we lack a clear understanding of why certain texts, though evident to human observers, elude scene text detectors. These complexity and rarity cases are grouped into this category, underscoring the need for further investigation into these atypical failures.

4.2 Pipeline of Constructing Benchmark

While these unresolved challenges have been identified, existing benchmarks either do not focus on them or only concentrate on one specific problem, limiting their utility in evaluating model performance across diverse and complex scenarios. To address this gap, we introduce the long-tailed Benchmark (LTB), which consists of 924 images, including a wide range of challenging text instances in various real-world scenes. LTB provides a more rigorous and comprehensive standard for assessing the performance of text detection models.

Our strategy for data collection prioritizes two main criteria: ensuring diversity in scenes and incorporating text instances that present predefined challenges. Based on these criteria, we choose six datasets, which are ICDAR2015 [Karatzas *et al.*, 2015], Total-Text [Ch'ng and Chan, 2017], ArT [Chng *et al.*, 2019], InverseText [Ye *et al.*, 2023], Occluded-RoadText, and NightTime-ArT [Yu *et al.*, 2023]. Detailed information on these benchmarks can be found in Appendix. According to the characteristics of different benchmarks, we adopt a dual strategy for data collection: one assisted by detectors and one purely manual. For ICDAR2015, Total-Text, and ArT, we obtain detection results of 4 well-finetuned detectors, as outlined in Table 2. We then develop an algorithm to retain undetected text instances. For the remaining three datasets, we manually collect challenging samples due to the absence of fine-tuned models demonstrating superior performance. Appendix elaborates our data processing strategy of LTB.

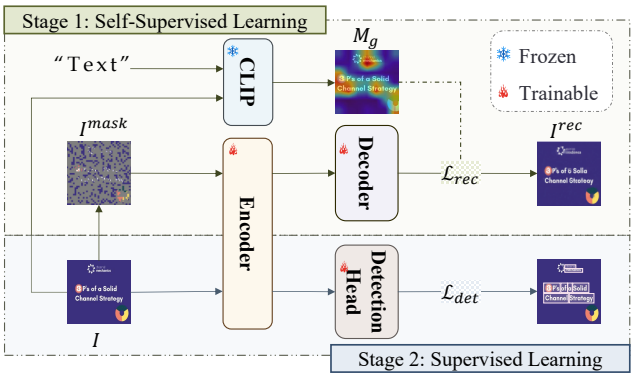

Figure 5: A detailed illustration of MAEDet architecture and training objectives. As indicated by its name, MAEDet is based on Mask Auto-Encoder (MAE), with the DBNet detection head adopted for text detection.

4.3 Baseline

Inspired by CLIP [Radford *et al.*, 2021] and MAE [He *et al.*, 2022], we adopt self-supervised learning to address the long-tailed problems from the data perspective, termed as MAEDet. As shown in the Figure 5, during the stage of SSL, given an image $I \in \mathbb{R}^{h \times w \times 3}$ and its randomly masked version I_{mask}, the model reconstructs the input as I^{rec}. To focus on the text region and model the text feature better, a guidance mask M_g is generated by the frozen CLIP. Specifically, M_g is the attention map, and M is the binary map with $\mathbf{T}$, which is the threshold for discriminating between the text region and background, and "Text" is the prompt of the text encoder of CLIP. Considering reconstruction bias between the text region and the background, we add an auxiliary branch to optimize MAEDet with the pixel-level balanced reconstruction

https://rrc.cvc.uab.es/?ch=29

	Method	Train Dataset	Test Dataset								
			ICDAR 2013	ICDAR 2015	TT	COCO	TextOCR	ArT	LSVT	MLT 2017	MLT 2019	Avg.
Pretrain	DBNet++	ST800K	66.6	22.8	39.5	27.5	11.1	26.7	12.1	56.0	52.3	35.0
	ABCNet v2	ST150K+COCO*+MLT19*	88.1	78.0	81.0	53.2	40.5	69.3	48.2	87.0	83.6	69.9
	DPText	ST150K+TT+MLT19	89.3	75.3	80.4	59.5	43.5	72.1	52.3	84.2	83.4	71.1
	SRFormer	ST150K+TT+MLT17	**90.6**	**81.5**	84.4	59.4	51.9	75.0	59.2	**90.1**	**86.8**	75.4
Finetune	DBNet	ICDAR2015	75.8	84.0	69.5	60.5	61.0	49.9	16.1	75.9	76.2	63.2
	DRRG	CTW1500	42.0	44.0	39.7	31.9	18.8	48.0	56.8	37.5	41.2	40.0
	FCENet	Total-Text	75.4	72.8	82.6	57.9	48.6	62.2	33.3	80.0	75.2	65.3
	LRANet	Total-Text	85.4	79.7	**88.9**	61.7	57.3	66.7	21.6	86.6	82.4	70.0
	PSENet	ICDAR2015	60.8	78.7	53.5	56.0	50.8	38.7	13.9	58.6	57.6	52.0
	TCM	ICDAR2015	81.6	88.0	79.0	65.9	67.0	57.6	18.2	82.6	82.8	69.2
Joint	DBNet++	Joint98K	87.1	69.3	84.4	59.5	58.3	78.9	73.4	83.3	77.1	74.6
	PANet	Joint98K	82.9	75.6	75.7	**64.5**	54.0	74.8	72.2	80.0	79.7	73.3
	DPText	Joint98K	88.2	75.9	81.9	57.3	**61.6**	**88.2**	**75.9**	87.5	80.3	**77.4**
SSL	MAEDet†	Joint98K	85.3	61.3	79.4	54.1	56.3	74.8	68.5	81.1	76.4	70.8
	MAEDet	Joint98K	87.1	67.5	82.8	58.7	60.6	78.4	71.8	83.3	80.3	74.5

Table 3: Comparison of MAEDet with other STD models across all nine test sets within Joint98K. F-measure (%) is reported. "$*$" denotes only a portion of the training images are used. "†" denotes model is trained without self-supervised weights. Abbreviations: TT = Total-Text, COCO = COCO-Text, ST800K = SynthText800K, ST150K = SynthText150K, SSL=self-supervised learning. **Bold** indicates the best result.

	Method	Intra-instance							Inter-instance		Background			Others	Overall	
		Blurred	Artistic	Glass	Single-Char	Inverse	Distorted	Delimited	Dense	Overlapped	Occluded	Low-Contrast	Complex-BG	Others	Hard	Norm
Pretrain	DBNet++	5.6	21.9	19.9	2.9	35.0	19.4	25.0	8.10	21.8	20.3	28.0	9.20	5.50	20.5	31.3
	ABCNet v2	13.6	21.0	19.1	7.8	23.1	9.90	13.8	17.9	15.1	26.6	45.4	14.2	7.90	27.1	46.7
	DPText	13.6	23.7	23.0	8.0	23.3	12.1	16.2	18.3	14.5	26.1	45.2	16.4	10.0	27.9	48.7
	SRFormer	22.9	28.2	24.6	15.3	23.7	18.2	16.0	25.6	19.7	33.5	46.8	21.9	7.80	33.2	53.3
Finetune	DBNet	27.4	26.5	27.1	2.71	43.9	27.5	13.6	36.2	23.7	31.0	50.2	11.1	8.25	34.8	59.4
	DRRG	7.57	17.8	30.1	4.81	58.1	25.6	40.5	5.04	21.4	27.6	34.6	21.4	10.2	24.9	35.1
	FCENet	22.4	41.3	**49.2**	11.9	75.6	**42.7**	22.8	29.5	34.3	32.3	60.7	36.0	25.2	41.1	65.0
	Mask-RCNN	24.5	17.9	18.5	6.15	30.4	13.0	19.4	27.6	26.4	32.6	42.7	7.46	9.04	31.4	53.6
	PSENet	17.7	15.3	15.3	3.24	16.7	14.9	12.7	27.1	21.9	25.7	35.3	1.94	3.74	25.0	48.5
	TCM	**35.0**	35.5	7.20	42.1	37.2	19.5	**62.5**	**39.9**	**36.9**	38.5	58.7	23.1	15.4	43.4	**68.4**
Joint	DBNet++	21.2	**44.2**	49.0	17.1	**76.1**	41.8	40.7	21.6	30.6	38.3	**64.3**	**42.2**	22.2	45.0	66.9
	PANet	23.6	34.2	37.3	19.2	60.5	33.9	30.7	23.9	20.5	38.9	61.7	23.2	5.50	42.1	65.1
	DPText	20.7	25.1	21.1	19.9	27.3	12.4	19.7	24.7	14.3	32.3	48.0	15.8	10.1	33.2	54.2
SSL	MAEDet†	19.6	41.1	38.4	17.0	66.0	37.4	44.2	21.7	20.4	38.1	58.6	23.4	**31.1**	41.2	63.5
	MAEDet	23.3	43.5	45.2	25.1	73.3	41.1	45.7	25.7	28.6	**43.9**	63.3	23.3	23.9	**45.7**	66.9

Table 4: Comparison of MAEDet with other STD models on LTB under different training strategies. F-measure (%) is reported. "†" denotes model is trained without self-supervised weights. "SSL" for self-supervised learning. **Bold** is the best and underline is second best.

loss $\mathcal{L}_{br}$, formulated as Equation (1), where α is the balance factor, $\mathbb{1}$ is the indicator function, and $\mathcal{L}_2$ is the L2 loss function. At the supervised learning stage, MAEDet is initialized from the pretrained backbone and optimized using detection loss $\mathcal{L}_{det}$.

$$\mathcal{L}_{br} = \sum_{i=1}^{H}\sum_{j=1}^{W} \alpha \mathbb{1}_{M_g > \mathbf{T}} \mathcal{L}_2\left(I_{ij} - I_{ij}^{\text{rec}}\right) + (1-\alpha)\mathbb{1}_{M_g \leq \mathbf{T}} \mathcal{L}_2\left(I_{ij} - I_{ij}^{\text{rec}}\right). \quad (1)$$

5 Experiments

In this section, we will first verify the performances of the representative detectors on mainstream benchmarks with the proposed JDL paradigm. Subsequently, the LTB benchmark is leveraged to evaluate further the long-tailed ability of these models, along with our baseline model MAEDet.

5.1 Experimental Setup

Implement Details. We use three different training strategies: 1) Pretrain: We utilize officially pretrained models for each detector, typically initialized with ImageNet pretraining weights; 2) Joint: Models are trained from scratch on the

# Exp.	Setup	LTB		Joint98K
		Hard	Norm	Avg.
Input Size				
1	480	34.3	53.5	63.8
2	640	41.0	62.9	71.3
3	800	**43.8**	**67.5**	**74.4**
Mask Threshold				
4	$\mathbf{T} = 0.1$	**44.8**	66.3	**74.2**
5	$\mathbf{T} = 0.2$	44.7	**66.4**	73.9
6	$\mathbf{T} = 0.3$	44.2	65.8	73.7
Optimization objective				
7	$\mathcal{L}_{rec}$	45.4	66.4	74.1
8	$\mathcal{L}_{rec} + \mathcal{L}_{br}$	**45.7**	**66.9**	**74.5**
Balance Factor				
9	$\alpha = 0.8$	42.0	63.5	71.6
10	$\alpha = 0.9$	**45.7**	**66.9**	**74.5**
11	$\alpha = 1.0$	44.7	66.4	73.9

Table 5: Ablation study results for MAEDet using Joint-Dataset Learning.

union of previous English word-level training sets, termed Joint98K, without loading any pretrained weights; 3) SSL: Our proposed MAEDet incorporate self-supervised learning using MARIO-LAION [Chen *et al.*, 2024] before training on Joint98K.

Evaluation Protocols. We employ F-measure [Karatzas *et al.*, 2015] as the standard evaluation metric, focusing solely on text instances labeled as "care" and disregarding "don't care" regions. For the LTB dataset, we report the F-measure for each of its 13 categories. Additionally, we use two overall metrics: *Hard*, which marks only challenging text instances as "care," and *Norm*, which retains the original ground truth annotations. These metrics provide complementary insights, with the Hard metric better differentiating model performance in challenging scenarios.

5.2 Main Results and Analysis

Table 3 presents a comparison between them. DBNet++ and DPText-DETR, when evaluated with the proposed JDL paradigm, achieve higher performance on more challenging datasets (COCO, ArT, LSVT, and TextOCR) than "Pretrain" settings. Additionally, the pretrained SRFormer exhibits excessively high performance on Total-Text, MLT2017, and MLT2019, likely due to data leakage, where its pertaining data includes some validation images from Joint98K.

Moreover, we also verify the effectiveness of JDL in reducing over-fitting in STD evaluation protocol between. As observed in Table 1, when DPText-DETR is fine-tuned on Total-Text, a notable performance gap of 15.1% emerges between its performance on ICDAR2015 and Total-Text. However, this discrepancy is significantly reduced to merely 6.0% when employing the JDL protocol. This substantial reduction in performance variance across different datasets substantiates the effectiveness of JDL in enhancing model generalization and alleviating the dataset-specific overfitting issue.

Table 4 presents a comparison of different training strategies on LTB. The results indicate that models trained on Joint98K achieve superior performance over their pretrained counterparts, suggesting that exposure to diverse, real-world data can positively influence model generalization under long-tailed conditions.

5.3 Ablation Studies

Exp. 1-3 in Table 5 demonstrate that appropriately increasing the input size for the LTB is beneficial for capturing challenging text instances. Mask threshold $\mathbf{T}$ in Equation (1) determines the proportion of text the backbone perceives during the self-supervised pretraining stage. Exp. 4-6 Table 5 show the best performance occurs when $\mathbf{T} = 0.1$. We hypothesize that a lower $\mathbf{T}$ allows more text pixels to contribute to the loss calculation. All experiments in this ablation study are conducted with $\alpha = 1.0$. Exp. 7-8 shows that after optimizing pretraining with $\mathcal{L}_{br}$, there is a significant improvement in performance from LTB to Joint98K, demonstrating that our auxiliary loss is effective in representing text features for scene text detection. The balance factor α controls the proportion of text and non-text regions perceived during the self-supervised learning stage. Exp. 9-11 demonstrate that MAEDet achieves the best performance when $\alpha = 0.9$, suggesting that some background pixels should be included in the comprehensive optimization objective. Therefore, we use $\alpha = 0.9$ as the default settings.

6 Conclusion

The paper identifies the fine-tuning and long-tailed problems as two significant factors contributing to the discrepancy between scene text detector performance on academic benchmarks and practical applications. Experiments demonstrate that models optimized for a single dataset often overfit specific domains but struggle to generalize to real-world scenarios. Given that Dataset-Specific Optimization (DSO) leads to such fine-tuning gap, we advocate Joint-Dataset Learning (JDL) for practical demands. Regarding the long-tailed problems, we define 13 categories of long-tailed challenges and introduce the Long-Tailed Benchmark (LTB), the first holistic benchmark to evaluate model's ability to detect different types of challenging texts, setting a new standard in scene text detection. Additionally, MAEDet is proposed as the baseline for our benchmark. We hope this work could provide valuable insights into improving scene text detectors and envision more efforts being devoted to practical approaches that effectively address long-tailed problems.

Acknowledgments

Supported by the National Natural Science Foundation of China (Grant NO 62376266 and 62406318), Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China, Beijing, China.

References

[Bu *et al.*, 2024] Qingwen Bu, Sungrae Park, Minsoo Khang, and Yichuan Cheng. SRFormer: Text detection transformer with incorporated segmentation and regression. In *AAAI*, volume 38, pages 855–863, 2024.

[Chen *et al.*, 2024] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *NeurIPS*, 36, 2024.

[Ch'ng and Chan, 2017] 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.

[Chng *et al.*, 2019] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *ICDAR*, pages 1571–1576. IEEE, 2019.

[Du *et al.*, 2022] Bo Du, Jian Ye, Jing Zhang, Juhua Liu, and Dacheng Tao. I3CL: Intra-and inter-instance collaborative learning for arbitrary-shaped scene text detection. *IJCV*, 130(8):1961–1977, 2022.

[Gupta *et al.*, 2016] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localization in natural images. In *CVPR*, pages 2315–2324, 2016.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, pages 2961–2969, 2017.

[He *et al.*, 2022] 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.

[Jaderberg *et al.*, 2016] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *International journal of computer vision*, 116:1–20, 2016.

[Jiang *et al.*, 2023] Qing Jiang, Jiapeng Wang, Dezhi Peng, Chongyu Liu, and Lianwen Jin. Revisiting scene text recognition: A data perspective. In *ICCV*, pages 20543–20554, 2023.

[Karatzas *et al.*, 2013] 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.

[Karatzas *et al.*, 2015] 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.

[Liao *et al.*, 2017] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *AAAI*, volume 31, 2017.

[Liao *et al.*, 2018] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE TIP*, 27(8):3676–3690, 2018.

[Liao *et al.*, 2020] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *AAAI*, volume 34, pages 11474–11481, 2020.

[Liu *et al.*, 2017] Yuliang Liu, Lianwen Jin, Shuaitao Zhang, and Sheng Zhang. Detecting curve text in the wild: New dataset and new solution. *arXiv*, 2017.

[Liu *et al.*, 2020] Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. ABCNet: Real-time scene text spotting with adaptive bezier-curve network. In *CVPR*, pages 9809–9818, 2020.

[Liu *et al.*, 2021] Yuliang Liu, Chunhua Shen, Lianwen Jin, Tong He, Peng Chen, Chongyu Liu, and Hao Chen. Abcnet v2: Adaptive bezier-curve network for real-time end-to-end text spotting. *IEEE TPAMI*, 44(11):8048–8064, 2021.

[Liu *et al.*, 2023] Ruijin Liu, Ning Lu, Dapeng Chen, Cheng Li, Zejian Yuan, and Wei Peng. Pbformer: Capturing complex scene text shape with polynomial band Transformer. In *ACM MM*, pages 2112–2120, 2023.

[Long *et al.*, 2018] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. Textsnake: A flexible representation for detecting text of arbitrary shapes. In *ECCV*, pages 20–36, 2018.

[Long *et al.*, 2022] 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.

[Lyu *et al.*, 2025] Jiahao Lyu, Wei Wang, Dongbao Yang, Jinwen Zhong, and Yu Zhou. Arbitrary reading order scene text spotter with local semantics guidance. In *AAAI*, volume 39, pages 5919–5927, 2025.

[Nayef *et al.*, 2017] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *ICDAR*, volume 1, pages 1454–1459. IEEE, 2017.

[Nayef *et al.*, 2019] 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. Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019. In *ICDAR*, pages 1582–1587. IEEE, 2019.

[Qin *et al.*, 2023] Xugong Qin, Pengyuan Lyu, Chengquan Zhang, Yu Zhou, Kun Yao, Peng Zhang, Hailun Lin, and Weiping Wang. Towards robust real-time scene text detection: From semantic to instance representation learning. In *ACM MM*, pages 2025–2034, 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack

Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Raisi *et al.*, 2021] Zobeir Raisi, Mohamed A Naiel, Georges Younes, Steven Wardell, and John S Zelek. Transformer-based text detection in the wild. In *CVPRW*, pages 3162–3171, 2021.

[Shen *et al.*, 2023] Huawen Shen, Xiang Gao, Jin Wei, Liang Qiao, Yu Zhou, Qiang Li, and Zhanzhan Cheng. Divide rows and conquer cells: Towards structure recognition for large tables. In *IJCAI*, pages 1369–1377, 2023.

[Shen *et al.*, 2024] Huawen Shen, Chang Liu, Gengluo Li, Xinlong Wang, Yu Zhou, Can Ma, and Xiangyang Ji. Falcon-ui: Understanding gui before following user instructions. *arXiv preprint arXiv:2412.09362*, 2024.

[Singh *et al.*, 2021] 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.

[Su *et al.*, 2024] Yuchen Su, Zhineng Chen, Zhiwen Shao, Yuning Du, Zhilong Ji, Jinfeng Bai, Yong Zhou, and Yu-Gang Jiang. Lranet: Towards accurate and efficient scene text detection with low-rank approximation network. In *AAAI*, volume 38, pages 4979–4987, 2024.

[Sun *et al.*, 2019] 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.

[Tang *et al.*, 2019] Jun Tang, Zhibo Yang, Yongpan Wang, Qi Zheng, Yongchao Xu, and Xiang Bai. Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping. *PR*, 96:106954, 2019.

[Tang *et al.*, 2022] Jingqun Tang, Wenqing Zhang, Hongye Liu, MingKun Yang, Bo Jiang, Guanglong Hu, and Xiang Bai. Few could be better than all: Feature sampling and grouping for scene text detection. In *CVPR*, pages 4563–4572, 2022.

[Tian *et al.*, 2022] Zichen Tian, Chuhui Xue, Jingyi Zhang, and Shijian Lu. Domain adaptive scene text detection via subcategorization. *arXiv*, 2022.

[Wang *et al.*, 2019] Wenhai Wang, Enze Xie, Xiaoge Song, Yuhang Zang, Wenjia Wang, Tong Lu, Gang Yu, and Chunhua Shen. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *ICCV*, pages 8440–8449, 2019.

[Wang *et al.*, 2020] Fangfang Wang, Yifeng Chen, Fei Wu, and Xi Li. TextRay: Contour-based geometric modeling for arbitrary-shaped scene text detection. In *ACM MM*, pages 111–119, 2020.

[Wang *et al.*, 2022] Wei Wang, Yu Zhou, Jiahao Lv, Dayan Wu, Guoqing Zhao, Ning Jiang, and Weipinng Wang. TPSNet: Reverse thinking of thin plate splines for arbitrary shape scene text representation. In *ACM MM*, pages 5014–5025, 2022.

[Xiao *et al.*, 2020] Shanyu Xiao, Liangrui Peng, Ruijie Yan, Keyu An, Gang Yao, and Jaesik Min. Sequential deformation for accurate scene text detection. In *ECCV*, pages 108–124. Springer, 2020.

[Ye *et al.*, 2020] Jian Ye, Zhe Chen, Juhua Liu, and Bo Du. Textfusenet: Scene text detection with richer fused features. In *IJCAI*, pages 516–522. International Joint Conferences on Artificial Intelligence Organization, 2020.

[Ye *et al.*, 2023] Maoyuan Ye, Jing Zhang, Shanshan Zhao, Juhua Liu, Bo Du, and Dacheng Tao. DPText-DETR: Towards better scene text detection with dynamic points in Transformer. In *AAAI*, volume 37, pages 3241–3249, 2023.

[Yu *et al.*, 2023] Wenwen Yu, Yuliang Liu, Wei Hua, Deqiang Jiang, Bo Ren, and Xiang Bai. Turning a CLIP model into a scene text detector. In *CVPR*, pages 6978–6988, 2023.

[Zeng *et al.*, 2023] Gangyan Zeng, Yuan Zhang, Yu Zhou, Bo Fang, Guoqing Zhao, Xin Wei, and Weiping Wang. Filling in the blank: Rationale-augmented prompt tuning for textvqa. In *ACM MM*, pages 1261–1272, 2023.

[Zhan *et al.*, 2019] Fangneng Zhan, Chuhui Xue, and Shijian Lu. Ga-dan: Geometry-aware domain adaptation network for scene text detection and recognition. In *ICCV*, pages 9105–9115, 2019.

[Zhang *et al.*, 2025] Yifei Zhang, Chang Liu, Jin Wei, Xiaomeng Yang, Yu Zhou, Can Ma, and Xiangyang Ji. Linguistics-aware masked image modeling for self-supervised scene text recognition. In *CVPR*, 2025.

[Zhou *et al.*, 2017] 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.

[Zhu *et al.*, 2021] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *ICLR*. OpenReview.net, 2021.