TextDS: Parameter-Efficient Representation Alignment for Scene Text Detection under Distribution Shifts

Boyuan Chen[1*], Zichen Dang[2*], Chuang Yang[2], Lap-Pui Chau[2], and Yi Wang[2†]

[1] School of Electrical Engineering, Xi'an Jiaotong University
[2] Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University
15956368@stu.xjtu.edu.cn, zichen.dang@connect.polyu.hk
{c1yang,lap-pui.chau,yi-eie.wang}@polyu.edu.hk

Abstract. In real-world deployments, scene text detectors inevitably face distribution shifts beyond the training distribution. Prior work often depends on large-scale scene-text pretraining, yet evaluation under cross-domain changes and real-world imaging degradations remains limited. We propose TextDS, an efficient framework for scene text detection under distribution shifts. First, we propose a data-efficient dual-encoder design with visual foundation models, eliminating the reliance on large-scale scene-text pretraining. Second, we introduce Step-wise LoRA adaptation (SWLoRA), which performs progressive low-rank refinement with a dynamic early-exit mechanism for effective feature adaptation. Third, we propose Common Subspace Fusion (CSF) to align and fuse the two branches in a shared subspace while retaining complementary, shift-robust information. Finally, we construct adverse-condition scene text detection datasets to address the gap in evaluating under imaging degradation. Experiments show that TextDS achieves competitive performance in scene text detection, demonstrating robustness across domains and adverse imaging conditions with only 4.9M trainable parameters. The code is publicly available at https://github.com/ZChenDang/TextDS

Keywords: Scene text detection · Distribution shifts · Adverse-condition scene

1 Introduction

Scene text detection aims to localize text instances in natural scenes, and serves as a foundation for end-to-end OCR and downstream understanding [24]. Reliable scene text detection enables a wide range of real-world applications, including instant translation and navigation [37], mobile document capture and payment [26], industrial inspection [8], and embodied robotic perception [31]. As

[*] Equal Contribution [†] Corresponding Author.

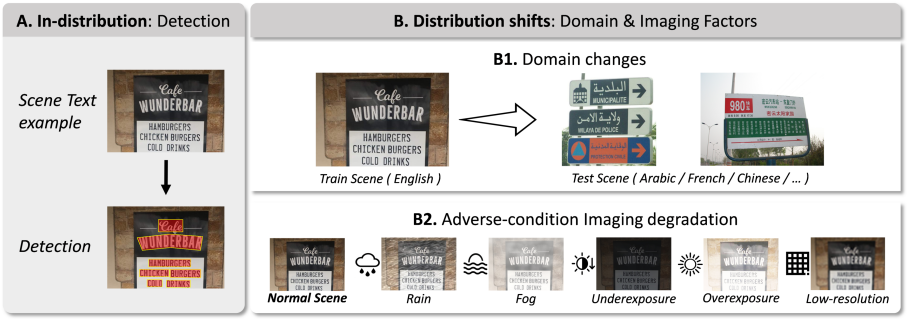

Fig. 1: Examples of distribution shifts in scene text detection, including domain changes and adverse-condition imaging degradation. The example of domain changes is transformation of text language from English to French, Arabic and Chinese. Imaging degradation includes low-resolution, rain, fog, underexposure and overexposure.

the scene text often carries dense, explicit semantic cues, the quality of detection directly sets the ceiling for subsequent recognition and parsing [18, 40, 41]. With increasing demand for large-scale, low-latency, and cross-device deployment, developing methods that are not only accurate but also stable and deployment-friendly is of significant scientific and practical importance [19, 39].

Current scene text detection methods can be categorized into two groups, including pixel aggregation approaches [16, 38, 49] and explicit structure modeling approaches [33, 34, 50]. Overall, pixel aggregation and explicit structure modeling represent two complementary paradigms for scene text detection, emphasizing dense prediction with instance aggregation and direct instance-level structure prediction, respectively.

However, existing methods are developed under relatively ideal acquisition conditions [7, 47], and rely heavily on large-scale pre-training on datasets such as SynthText-800k or SynthText-150k [9] to improve in-distribution performance, leaving evaluation under distribution shifts insufficiently studied and rarely optimized. In real-world scene deployments, scene text detectors encounter distribution shifts relative to the training distribution [20], driven by two types: **cross-domain changes in data sources** and **imaging degradations introduced during capture and transmission** [46], which is illustrated in Fig. 1. The former type of domain shift involves changes in distribution caused by differences in data sources, such as variations in shooting scenes, font languages, and font styles, which result in the actual application data distribution deviating from the training distribution [44]. For the latter type, real imaging degradations are pervasive and physically grounded [11]. For instance, rain and fog reduce contrast through scattering, low-light and underexposure compress dynamic range and increase noise [1, 4], strong illumination or exposure failures cause overexposure and highlight saturation that erase stroke and boundary details [51], and long-range capture, digital zoom, and compressed transmission produce low-resolution inputs with lost fine textures [32]. Such shifts simulta-

neously undermine pixel separability and geometric cues, thereby destabilizing thresholding and region growing in pixel-aggregation methods and weakening boundary localization and component association in structure-modeling methods. Despite this, specific datasets and systematic evaluations for these adverse conditions remain scarce, hindering robust assessment under unified distribution-shift settings.

To address the aforementioned issues, we propose TextDS, an efficient scene text detection network designed for distribution-shift settings. TextDS is to achieve both data-efficient and parameter-efficient detection with only a small number of trainable parameters, without relying on large-scale scene-text-specific pretraining, while improving cross-domain generalization and enabling robust evaluation under degraded imaging conditions. The main contributions of this paper are summarized as follows:

- **Dual-branch feature extraction and fusion.** We propose a dual-branch architecture on complementary visual foundation models for scene text feature extraction, and propose Common Subspace Fusion (CSF) to align and fuse the two branches in a shared subspace, avoiding reliance on large-scale scene-text-specific pretraining, with a small trainable parameter count.
- **Dynamic step-wise low-rank adaptation.** We propose Step-wise Low-rank Adaptation (SWLoRA) with a cosine-similarity based dynamic early-exit mechanism, enabling progressive refinement and adaptive stopping of the step-wise updates across samples.
- **Extensive evaluation across domains and degradations.** We validate our method with extensive experiments across multiple domains, and complement standard evaluation with proposed adverse-condition datasets to facilitate robustness analysis under diverse degradations.

2 Related Work

2.1 Scene Text Detection Methods

We briefly review representative scene text detection approaches from two complementary perspectives: pixel aggregation and explicit structure modeling.

Pixel aggregation approaches are centered on dense prediction, such as probability maps, kernel or center regions, and direction or distance fields, and then aggregate pixels into text instances via binarization, connected components, and region growing. Liao *et al.* [16] proposed DB-Net, which makes binarization differentiable and learns adaptive thresholds to reduce reliance on handcrafted thresholds and complicated aggregation rules. Building on the idea of improving supervision and reconstruction robustness, Zhang *et al.* [49] proposed TextPMs to address uncertainty introduced by coarse-grained annotations by using a set of probability maps to model text-pixel probability distributions and then reconstructing instances via iterative prediction and region growing. Following this direction, Wang *et al.* [38] proposed S3INet, which strengthens multi-scale semantics and foreground prominence via semantic spatial interaction on single-level features, and suppress text-like backgrounds, thereby reducing false positives.

In contrast, explicit structure modeling approaches treat instance structure as the direct prediction target, and incorporate boundary constraints or component relations into end-to-end learning to reduce dependence on heuristic aggregation. Along this line, Zhang *et al*. [50] proposed TextBPN, which refines contour details via boundary proposals and iterative deformation. From a compact shape representation perspective, Su *et al*. [34] proposed LRANet, which learns data-driven shape bases via low-rank approximation and reconstructs text contours using a small number of coefficients to balance accuracy and efficiency. Moving from contour parameterization to component relation modeling, Su *et al*. [33] proposed ERRNet, which decomposes text into an ordered sequence of components and explicitly models inter-component relations from a tracking perspective, enabling component-level inference without post-processing. Overall, while pixel aggregation methods achieve instance formation through dense predictions and subsequent grouping, explicit structure modeling methods aim to predict instance geometry or component relations directly.

In addition, in recent years, end-to-end OCR based on large language models has also been introduced into the scene text detection and recognition process, gradually shifting from the detection paradigm to stronger general visual understanding capabilities. Qwen-OCR [3], Deepseek-OCR [42], and PP-OCR [14] provide more engineering-oriented and readily deployable solutions. Overall, the large language model-based approach has enhanced the semantic understanding ability for complex scene texts, while the scene text detection methods still have competitiveness in terms of real-time performance and controllability. Both are also showing a trend of integration.

2.2 Foundation Models: Representation, Segmentation, and Efficient Adaptation

Vision foundation models offer strong general-purpose priors that can be transferred to downstream tasks with limited task-specific data. In particular, DINO [5] learns Vision Transformer representations through self-distillation without manual annotations, producing features with strong semantic aggregation that transfer well to diverse tasks [2, 12, 15]. Building on this line, DINOv2 [27] and DINOv3 [30] further advance the pretraining strategy and scaling to yield even stronger transferability. For scene text detection, prior practice often relies on large-scale text-specific synthetic pretraining such as SynthText [9]. In contrast, DINOv3 provides an alternative path where general self-supervised weights can be adapted to text detection with lightweight task-specific components.

Alongside strong representations, promptable segmentation models further broaden the applicability of foundation models. SAM [13] and SAM2 [28] decouple image encoding from mask prediction and generate class-agnostic masks conditioned on prompts (e.g., points and boxes), enabling reuse of dense representations across tasks. SAM-style models have been adopted in domains beyond natural images, including medical imaging [6, 23], remote sensing [21, 45], and video tracking [43]. However, their systematic exploration in scene text detection remains limited.

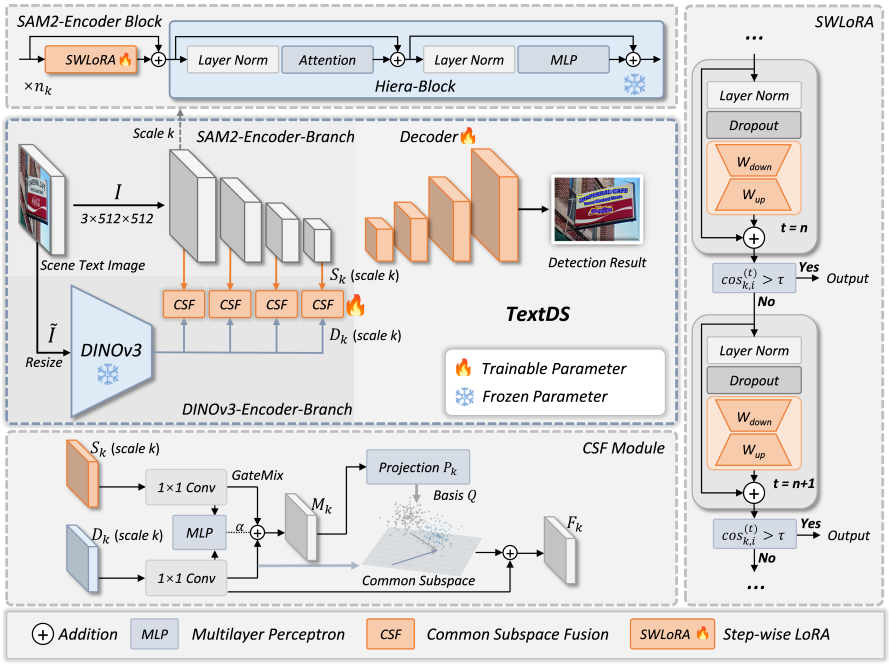

Fig. 2: Overall structure of TextDS. The input scene text image is processed through the SAM2-Encoder-Branch and the DINOv3-Encoder-Branch, and the dual-branch encoded features of multiple scales are fused through the Common Subspace Fusion (CSF) module, and each SAM2 Encoder Block uses Step-wise LoRA (SWLoRA) structure for fine-tuning.

To make adapting large foundation models feasible under limited compute and data, LoRA [10] injects low-rank trainable updates into pretrained linear layers while freezing the backbone, enabling effective transfer with a small number of additional parameters and reduced training cost [22]. Recent variants such as AdaLoRA [48] and DoRA [17] further improve adaptation quality while keeping the updates lightweight, yet LoRA-style designs tailored to the specific challenges of scene text detection remain underexplored.

3 Method

3.1 Overview of TextDS Architecture

Our proposed TextDS follows an overall approach of constructing a dual-encoder using data-efficient representations of the source (SAM2 and DINOv3) that do not rely on scene text for large-scale pre-training. Then, through a parameter-efficient adaptation and fusion mechanism, the cross-domain robust general representation is transformed into the discriminative features required for scene

text detection. Specifically, the SAM2 encoding branch provides strong structure and multi-scale priors, while the DINOv3 encoding branch provides robust semantic representation. We propose Step-wise LoRA (SWLoRA) at the block-level input of the SAM2 backbone to complete task adaptation, and propose Common Subspace Fusion (CSF) to fuse the two branches within the shared subspace, while retaining the domain-robust information of DINOv3 in the orthogonal complement space. Finally, we output the text probability map with a lightweight decoder. The overall structure is shown in Fig. 2.

3.2 SAM2-DINOv3 Dual-Encoder Architecture

To achieve robust scene text detection under distribution shifts while maintaining a data-efficient representation, we adopt a dual-encoder architecture composed of SAM2-Hiera-L and DINOv3 ViT-L/16. Given an input image

$$I \in \mathbb{R}^{B\times 3\times H\times W}, \tag{1}$$

where B denotes the batch size and $H \times W$ is the input resolution, the SAM2 encoding branch produces a four-level feature pyramid capturing strong structural priors, while the DINOv3 branch provides a high-level, domain-robust semantic representation. We then align DINOv3 features to each SAM2 scale for subsequent cross-branch fusion and decoding.

SAM2-Hiera-L for multi-scale structural encoding. We keep only the SAM2 image encoder trunk (removing the prompt encoder, mask decoder, and memory-related components), as shown in Fig. 2, and use it as a frozen structural feature extractor. The SAM2 trunk yields four-scale features

$$\{S_1, S_2, S_3, S_4\} = E_{\text{sam2}}(I), \qquad S_k \in \mathbb{R}^{B\times C_k\times H_k\times W_k},\ k \in \{1,2,3,4\}, \tag{2}$$

where $E_{\text{sam2}}(\cdot)$ denotes the SAM2-Hiera-L trunk, S_k is the output at the k-th scale, and C_k, H_k, W_k are its channel dimension and spatial resolution, respectively. In our implementation, the channel dimensions are fixed to

$$(C_1, C_2, C_3, C_4) = (144,\ 288,\ 576,\ 1152). \tag{3}$$

These features form a pyramid from high to low resolution, following the hierarchical downsampling pattern. For a consistent description of hierarchical computation, we name the four scales as Hiera-S1/S2/S3/S4 and denote the set of Hiera blocks in the k-th scale by

$$\mathcal{B}_k = \{b_{k,1}, \ldots, b_{k,n_k}\}, \qquad k \in \{1,2,3,4\}, \tag{4}$$

where $b_{k,i}$ is the i-th block at scale k and n_k is the number of blocks in that scale, and the total number of blocks equals $\sum_{k=1}^{4} n_k$. In our model, each trunk block is wrapped by a parameter-efficient adapter (introduced in the next subsection) while the original SAM2 parameters remain frozen, enabling task adaptation with minimal trainable parameters.

DINOv3 ViT-L/16 semantic extraction and multi-scale alignment. For the DINOv3 encoding branch, we resize the input to a fixed resolution 448×448:

$$\tilde{I} = \text{Resize}(I, 448, 448). \tag{5}$$

We take the last output feature of DINOv3 (the last element returned by a features_only interface), denoted as

$$D = E_{\text{DINO}}^{(\text{last})}(\tilde{I}), \qquad D \in \mathbb{R}^{B\times 1024\times 28\times 28}. \tag{6}$$

Here, $E_{\text{DINO}}^{(\text{last})}$ represents the final semantic representation of ViT-L/16. The spatial size 28×28 comes from the patch size of 16 (i.e., $448/16 = 28$), and the channel dimension is 1024.

Since D is a single-scale feature map, we map and align it to each SAM2 scale via four 1×1 convolutions followed by bilinear resizing:

$$D_k = \text{Resize}(Align_k(D), (H_k, W_k)), Align_k : \mathbb{R}^{1024} \to \mathbb{R}^{C_k}. \tag{7}$$

This produces a scale-aligned pair (S_k, D_k) at each sampling level, where $C_k \in \{144, 288, 576, 1152\}$. The aligned multi-scale features are then fed into our fusion module and lightweight decoder for scene text prediction.

3.3 Step-wise LoRA Module

Scene text detection exhibits scale-dependent requirements: high resolution stage is critical for stroke boundaries and thin structures, middle stages govern instance connectivity and deformation consistency, and low-resolution stages mainly contribute global layout priors and background suppression. Following our SAM2-Hiera hierarchy defined in Section 3.2, we denote the Hiera blocks at stage k as shown in Eq. (4). We insert a step-wise LoRA module (SWLoRA) before each frozen Hiera block to perform a multi-step, low-rank residual refinement of the incoming token features. SWLoRA further employs a cosine-similarity criterion to dynamically stop the refinement once the representation stabilizes, allocating extra computation only to hard samples and hard stages.

For any block $b_{k,i} \in \mathcal{B}_k$, let its frozen mapping be $f_{k,i}(\cdot)$. With SWLoRA, the block output is computed as

$$\mathbf{y}_{k,i} = f_{k,i}\big(\text{SWLoRA}_{k,i}(\mathbf{x}_{k,i})\big), \tag{8}$$

where $\mathbf{x}_{k,i}$ and $\mathbf{y}_{k,i}$ are the input and output token representations of the block $b_{k,i}$, respectively.

Step-wise low-rank residual refinement. Instead of applying a single static LoRA update, SWLoRA performs up to T refinement steps. Let $\mathbf{x}_{k,i}^{(0)} = \mathbf{x}_{k,i}$. The refinement process can be represented as

$$\mathbf{x}_{k,i}^{(t+1)} = \mathbf{x}_{k,i}^{(t)} + \gamma_{k,i,t}\, \Delta_{k,i}\Big(\mathbf{x}_{k,i}^{(t)}\Big), \qquad t = 0, \ldots, T-1, \tag{9}$$

where $\gamma_{k,i,t}$ is a learnable step scale, and $\Delta_{k,i}(\cdot)$ is a low-rank residual mapping:

$$\Delta_{k,i}(\mathbf{x}) = \frac{\alpha}{r}\,\mathbf{W}_{\text{up}}^{(k,i)}\Big(\mathbf{W}_{\text{down}}^{(k,i)}\big(\text{Dropout}(\text{LN}(\mathbf{x}))\big)\Big). \tag{10}$$

Here $\text{LN}(\cdot)$ is layer normalization, r is the LoRA rank, α is the LoRA scaling factor, and $\mathbf{W}_{\text{down}}^{(k,i)}, \mathbf{W}_{\text{up}}^{(k,i)}$ are the down-/up-projection matrices. After refinement, SWLoRA outputs $\mathbf{x}_{k,i}^{(T^*)}$ to the frozen block, where $T^* \le T$ is the number of the executed steps.

Cosine-similarity based dynamic early-exit. Cosine-similarity based dynamic early-exit mechanism adaptively determines the effective refinement depth T^* for each sample and each block. The key motivation is that the required refinement depth is not constant: it varies with input difficulty and with the stage/block characteristics. In experiments, we set the maximum number of refinement steps to $T = 5$.

Specifically, after obtaining $\mathbf{x}_{k,i}^{(t+1)}$, we compute the cosine similarity between consecutive states:

$$\cos_{k,i}^{(t)} = \frac{\left\langle \text{vec}\left(\mathbf{x}_{k,i}^{(t+1)}\right),\ \text{vec}\left(\mathbf{x}_{k,i}^{(t)}\right)\right\rangle}{\left\|\text{vec}\left(\mathbf{x}_{k,i}^{(t+1)}\right)\right\|_2 \left\|\text{vec}\left(\mathbf{x}_{k,i}^{(t)}\right)\right\|_2 + \varepsilon}, \tag{11}$$

where $\text{vec}(\cdot)$ flattens the token and channel dimensions and ε is a small constant for numerical stability. A high $\cos_{k,i}^{(t)}$ indicates that the refinement update becomes directionally stable, suggesting that the current representation has already absorbed the necessary task-specific correction at this block.

Given a minimum-step constraint $T_{\min}$=1, we stop refinement early if

$$t + 1 \ge T_{\min} \quad \text{and} \quad \cos_{k,i}^{(t)} > \tau, \tag{12}$$

where τ is the exit threshold. Once Eq. (12) is satisfied, we set $T^* = t + 1$ and output $\mathbf{x}_{k,i}^{(T^*)}$ to the frozen block, otherwise, the refinement continues until reaching the maximum step $T = 5$.

This dynamic mechanism encourages SWLoRA to allocate refinement depth where it is truly needed, especially for hard samples or blocks requiring stronger correction, and to stop once the representation stabilizes, leading to more effective feature adaptation across samples and blocks.

3.4 Common Subspace Fusion (CSF)

Given the scale-aligned feature pairs $\{(S_k, D_k)\}_{k=1}^{4}$, we fuse them to combine structural cues from SAM2 encoding branch with domain-robust semantics from the DINOv3 branch.

Channel alignment and gated mixing. We first project both branches to a common channel dimension C:

$$\hat{S}_k = \phi_s^{(k)}(S_k), \qquad \hat{D}_k = \phi_d^{(k)}(D_k), \qquad \hat{S}_k, \hat{D}_k \in \mathbb{R}^{B\times C\times H_k\times W_k}, \tag{13}$$

where $\phi_s^{(k)}$ and $\phi_d^{(k)}$ are lightweight 1×1 convolutions. We then compute an input-adaptive gate α_k from global statistics and form an intermediate mixture:

$$\alpha_k = \sigma\Big(\mathrm{MLP}\big([\mathrm{GAP}(\hat{S}_k); \mathrm{GAP}(\hat{D}_k)]\big)\Big), \qquad M_k = \alpha_k \odot \hat{S}_k + (1-\alpha_k)\odot \hat{D}_k, \tag{14}$$

where σ represents the Sigmoid function, α_k is a scalar or channel-wise weight, $\odot$ denotes element-wise multiplication, and $\mathrm{GAP}(\cdot)$ is global average pooling.

Common subspace estimation. Let $N_k = H_k W_k$ and denote by $X_{S_k}, X_{D_k} \in \mathbb{R}^{B\times C\times N_k}$ the spatially flattened, mean-centered features of $\hat{S}_k$ and $\hat{D}_k$, respectively. We estimate per-sample covariances and construct a second-order proxy for shared variation:

$$\Sigma_{S_k} = \frac{1}{N_k} X_{S_k} X_{S_k}^\top + \varepsilon \mathbf{I}, \quad \Sigma_{D_k} = \frac{1}{N_k} X_{D_k} X_{D_k}^\top + \varepsilon \mathbf{I}, \tag{15}$$

$$A_k = \Sigma_{S_k} \Sigma_{D_k}, \tag{16}$$

where $\varepsilon\mathbf{I}$ is added for numerical stability. We approximate the top-r eigenspace (r=16) of A_k and obtain an orthonormal basis $Q_k \in \mathbb{R}^{B\times C\times r}$.

Fusion in the common subspace with DINO complement preservation. Given Q_k, we define the projection onto the estimated common subspace:

$$\mathcal{P}_k(X) = Q_k(Q_k^\top X), \tag{17}$$

applied to flattened features and reshaped back to $\mathbb{R}^{B\times C\times H_k\times W_k}$. We restrict cross-branch interaction to the projected component, while explicitly preserving DINOv3's orthogonal complement:

$$F_k = \mathcal{P}_k(M_k) \; + \; \big(\hat{D}_k - \mathcal{P}_k(\hat{D}_k)\big), \qquad F_k \in \mathbb{R}^{B\times C\times H_k\times W_k}. \tag{18}$$

The first term enforces fusion only on the statistically shared subspace, whereas the second term prevents the fusion operator from implicitly removing DINOv3-specific components, and feed $\{F_k\}_{k=1}^4$ to the decoder for multi-scale prediction.

4 Experiments

4.1 Datasets

Normal-condition datasets. CTW-1500 [47] is a benchmark focusing on curved text with polygon-based annotations for long and irregular text lines.

Following the standard split, it contains 1000 training images and 500 testing images. **Total-Text [7]** covers multi-oriented and curved text in natural scenes and is widely used to evaluate localization under diverse layouts. It provides 1255 training images and 300 testing images in the commonly adopted split. **MLT [25]** (Multi-Lingual Text) targets multi-lingual, multi-script scene text with substantial appearance variation. Here we utilize 7200 training set images and 1800 images for testing.

Adverse-condition datasets. To systematically evaluate robustness under real-world imaging degradations, we construct degraded variants of CTW-1500, Total-Text, and MLT. Specifically, we create five subsets of **Rain**, **Fog**, **Underexposure**, **Overexposure**, and **Low-resolution** by applying corresponding degradation processes to original images while keeping ground-truth annotations unchanged. Detailed examples are provided in the supplementary material. We also conducted zero-shot experiments on low-light scenes under real nighttime conditions, with details in the supplementary materials.

4.2 Implementation Details

We implement our proposed TextDS in PyTorch and initialize it with two pretrained backbones: SAM2-Hiera-L and DINOv3 ViT-L/16. During training, images and corresponding masks are resized to 512×512. All experiments are conducted on a single NVIDIA RTX 4090 GPU. We train for 50 epochs with a batch size of 4 using the AdamW optimizer, with an initial learning rate of 0.001 and weight decay of 5×10^{-4}. The learning rate is scheduled with cosine annealing and a minimum learning rate of 1×10^{-7}. We optimize a structure loss function composed of an equal-weighted binary cross-entropy term and an IoU term to emphasize boundary regions.

4.3 Evaluation Metrics

We evaluate detection performance using Precision (P), Recall (R), and F-measure (F). Precision measures the fraction of predicted positives that are correct, while recall measures the fraction of ground-truth positives that are successfully detected. Considering both precision and recall comprehensively, we further use F-measure as the comprehensive indicator.

4.4 Comparison with State-of-the-art Methods

Our experimental results on the CTW-1500, Total-Text and MLT datasets under normal conditions without distribution shifts are shown in Tab. 1. The methods used for comparison include those from the representative approaches in recent years, including DB-Net [16], TextBPN [50], TextPMs [49], TextDCT [35], CT-Net [29], LRANet [34], CB-Net [52], ERRNet [33], and S3INet [38]. The TextDS

Table 1: Comparison of scene text detection accuracy and model complexity. Accuracy metrics include Precision (P), Recall (R), and F-measure (F) (%). Model complexity includes whether SynthText pre-training (Synth) is required, the number of trainable parameters (Param.), and FPS (uniformly measured on the same RTX 4090 GPU).

Method	Published	Synth	Param. (M)	CTW-1500			Total-Text			MLT			FPS
				P	R	F	P	R	F	P	R	F	
DB-Net [16]	AAAI 2020	✓	28.0	86.9	80.2	83.4	87.1	82.5	84.7	87.8	83.5	85.6	35.0
TextBPN [50]	ICCV 2021	✓	38.7	86.5	83.6	85.0	90.7	85.2	87.9	90.3	84.6	87.4	22.6
TextPMs [49]	TPAMI 2022	✓	36.4	87.8	83.8	85.7	90.6	86.5	88.5	88.1	83.6	85.8	16.8
TextDCT [35]	TMM 2023	✓	-	85.0	85.3	85.1	87.2	82.7	84.9	-	-	-	-
CT-Net [29]	TCSVT 2023	✓	-	87.9	82.7	85.2	90.8	85.0	87.8	-	-	-	-
LRANet [34]	AAAI 2024	✓	34.3	89.4	85.5	87.4	88.9	86.6	87.7	92.7	84.4	88.3	38.1
CB-Net [52]	IJCV 2024	✓	12.3	89.3	82.9	86.0	90.1	82.5	86.1	**92.9**	85.2	88.9	32.9
ERRNet [33]	AAAI 2025	✓	-	88.1	85.3	86.7	90.1	86.1	88.1	-	-	-	-
S3INet [38]	TNNLS 2025	✓	38.0	89.2	83.0	86.0	**91.2**	86.2	88.7	92.8	86.8	89.7	37.3
TextDS	-	✗	**4.9**	**91.6**	**89.6**	**90.6**	89.7	**88.4**	**89.1**	92.4	**92.0**	**92.2**	**44.1**

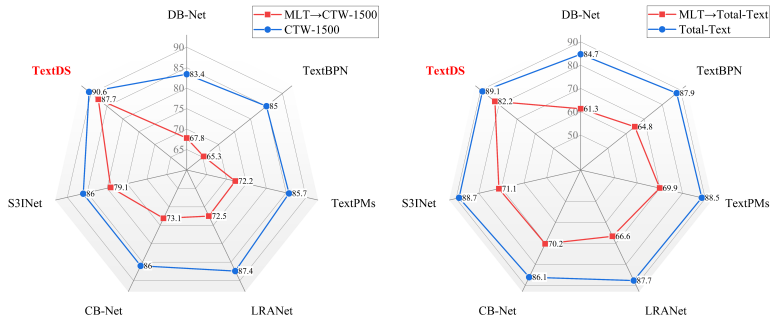

Fig. 3: The results of domain generalization from the MLT dataset to CTW-1500 and Total-Text are compared between TextDS and the comparison methods. The blue circles represent the F-measure of the model on CTW-1500 and Total-Text by itself, while the red circles represent the F-measure of the model when generalizing from MLT dataset to CTW-1500 and Total-Text.

we proposed significantly outperforms the comparison methods in terms of detection performance in Precision, Recall, and F-measure. Especially, it is notably superior to the methods developed in recent two years (2024 and 2025), indicating the outstanding performance of TextDS on the in-distribution cases.

4.5 Comparison of Model Complexity

While maintaining accuracy beyond existing models, the TextDS we propose, due to its Data- and Parameter-efficient structural design, requires a detailed analysis and comparison of its computational complexity. Model complexity comparison includes whether SynthText [9] pre-training is required, the number of trainable parameters, and the Frames Per Second (FPS) value. The specific results are shown in Tab. 1. To ensure rigor, results that were not code-open-sourced and not

Table 2: Experimental results of our proposed degraded scene text image datasets under adverse conditions. Imaging conditions include weather (rain and fog), exposure (underexposure and overexposure), and low-resolution (256×256 and 128×128).

Imaging Condition		CTW-1500			Total-Text			MLT		
		P	R	F	P	R	F	P	R	F
Normal		91.6	89.6	90.6	89.7	88.4	89.1	92.4	92.0	92.2
Weather	Rain	90.2	88.3	89.2	87.3	88.2	87.7	92.4	90.9	91.6
	Fog	91.7	88.0	89.8	88.4	88.2	88.3	92.5	91.4	91.9
Exposure	Underexposure	91.3	88.9	90.1	87.6	88.0	87.8	92.0	92.2	92.1
	Overexposure	89.9	90.0	89.9	88.0	87.9	87.9	91.9	91.7	91.8
Low-resolution	256×256	90.2	89.3	89.8	88.7	87.9	88.3	90.8	91.1	90.9
	128×128	88.4	89.4	88.9	87.0	89.2	88.0	86.9	89.9	88.3

Table 3: Ablation experimental results of the key components on F-measure (%).

Dual-Encoding	SWLoRA	CSF	CTW-1500	Total-Text	MLT
			82.2	85.1	86.4
✓			85.3	86.0	88.1
✓	✓		87.9	88.3	90.3
✓		✓	88.2	87.8	91.0
✓	✓	✓	**90.6**	**89.1**	**92.2**

reported in the past paper are indicated by (-). Compared with the comparison methods, the characteristics of TextDS make it no need to pre-train with the specific large-scale SynthText dataset for scene text detection. On this basis, TextDS only requires 4.9M parameters, significantly lower than the comparison methods. Moreover, while maintaining the same hardware configuration, TextDS achieves the highest 44.1 FPS value, demonstrating the efficiency of our design and providing a new approach suitable for practical deployment.

4.6 Experimental Results of Domain Generalization

The domain generalization results of TextDS and the comparison methods from the MLT dataset to CTW-1500 and Total-Text are shown in Fig. 3. The comparison methods include DB-Net [16], TextBPN [50], TextPMs [49], LRANet [34], CB-Net [52], and S3INet [38]. The blue circles represent the in-domain F-measure of the models on CTW-1500 and Total-Text, while the red circles represent the F-measure of the models when generalizing from MLT dataset to CTW-1500 and Total-Text. It can be intuitively observed that TextDS, when performing the best in domain-specific data performance, also exhibits significantly superior performance in domain generalization of different scene text detection datasets.

4.7 Experimental Results of Adverse-condition Datasets

The experimental results of the degraded dataset under adverse conditions are shown in Tab. 2, and the visualization is of degradation examples and detection results is shown in Fig. 4. Based on the CTW-1500, Total-Text and MLT

Fig. 4: Visualization of the detailed scene text image degradation and detection examples for TextDS. For the detection samples, the red color represents the Ground-Truth, while the blue color represents the text region results detected by the TextDS, which maintains performance under adverse imaging conditions.

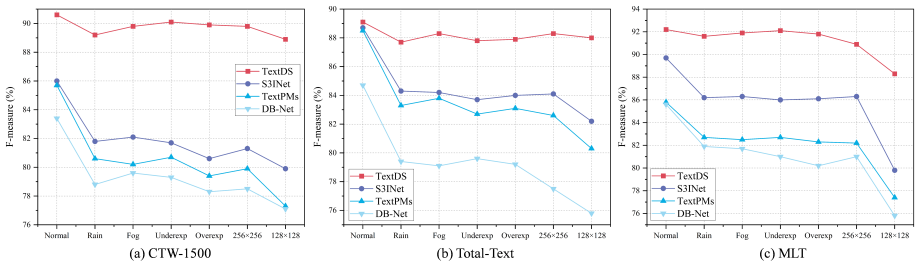

Fig. 5: Performance comparison of TextDS and the comparison methods on the scene text datasets in adverse-condition scenarios, including the F-measure of TextDS, S3INet, TextPMs and DBNet under Normal and degraded imaging conditions.

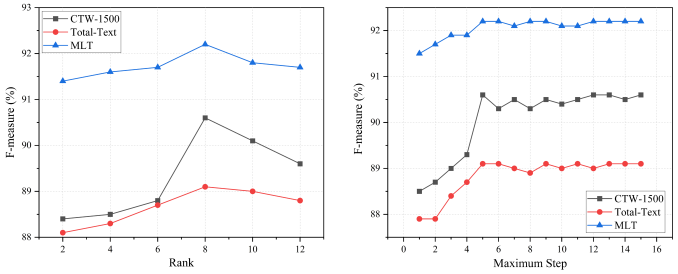

Fig. 6: Variation of the F-measure with the selection of Rank and Maximum Step, using the SWLoRA structure to fine-tune CTW-1500, Total-Text, and MLT datasets, where Rank = 8 and Maximum Step = 5 are adopted as the default setting.

datasets, we processed and obtained their respective datasets of adverse imaging conditions, including weather (rain and fog), exposure (underexposure and overexposure) and low-resolution (256×256 and 128×128). It can be seen that TextDS has demonstrated robustness when dealing with these degraded datasets. This part of the experiment fills the gap in the datasets for scene text detection under adverse imaging conditions. Fig. 5 compares the performance of TextDS with several representative baseline methods on the adverse-condition scene text dataset, reporting the F-measure of TextDS, S3INet [38], TextPMs [49], and DB-Net [16] under the normal setting and various degraded imaging conditions. As can be clearly observed, under degraded conditions, TextDS achieves consistently higher overall performance than the competing methods and exhibits a clearly smaller drop relative to the F-measure of normal setting, demonstrating its strong robustness to adverse imaging conditions.

4.8 Ablation Studies and Discussion

Ablation experiments of Key components. The ablation experimental results of TextDS are shown in Tab. 3, using F-measure as the evaluation metric. The key components include the Dual-Encoding, SWLoRA and CSF modules. The ablation experiments further demonstrated the effectiveness of the proposed separate and combined use of the structures. Detailed ablation experiments for datasets under adverse conditions are presented in the supplementary materials.

Selection of Rank and Step. Using the SWLoRA structure for fine-tuning, we further study how the final performance depends on two key hyper-parameters: the LoRA rank and the maximum step in SWLoRA. As shown in Fig. 6, increasing the Rank or Maximum Step generally improves the fitting ability at the beginning, but the gains quickly saturate and may become unstable when the adaptation capacity is overly large. Considering the overall trend across the three datasets, as well as parameter efficiency and convergence stability, we set Rank $= 8$ and Maximum Step $= 5$ as the default configuration for all experiments. Under this setting, the proposed early-exit mechanism further reduces unnecessary refinement computation, with average executed steps of 3.652, 2.205, and 3.476 out of 5 on CTW-1500, Total-Text, and MLT, respectively, corresponding to reductions of 27.0%, 55.9%, and 30.5%. These results show that SWLoRA can adaptively stop the refinement process once the representation becomes sufficiently stable, achieving a favorable balance between accuracy, stability, and computational efficiency.

Comparison with LLM-based and end-to-end OCR methods. We further conducted a visual qualitative comparison between proposed TextDS and representative LLM-based and end-to-end OCR methods, including Qwen-OCR [3], DeepSeek-OCR [42], and PP-OCR [14], as shown in Fig. 7. The LLM-based methods have strong capabilities in semantic understanding [36], but they have weaknesses in pixel-level geometric positioning of complex deformed texts.

Fig. 7: Visual comparison with LLM-based and end-to-end OCR methods in scene text detection, including Qwen-OCR, DeepSeek-OCR, and PP-OCR.

TextDS, as a front-end detector, due to its extremely strong structural prior, is highly complementary to the existing LLM-VLM models and can provide higher-quality text extraction regions for large language models. More comparison details and examples can be found in the supplementary material.

5 Conclusion

This paper addresses the practical challenge that scene text detectors face distribution shifts and real-world imaging degradations at deployment. We propose TextDS, an efficient scene text detection network that uses a data-efficient dual-encoder strategy combining SAM2 and DINOv3 to capture complementary text-related representations without large-scale scene-text-specific pretraining, and introduces Step-wise LoRA Adaptation and Common Subspace Fusion for parameter-efficient fine-tuning and alignment of the two encoding branches. In addition, we build adverse-condition scene text detection subsets to enable systematic evaluation under imaging degradations. Extensive experiments show that TextDS achieves competitive detection performance while improving robustness in domain generalization and adverse imaging conditions.

Acknowledgements

The research work described in this paper was conducted in the JC STEM Lab of Machine Learning and Computer Vision funded by The Hong Kong Jockey Club Charities Trust. This research received partially support from the Global STEM Professorship Scheme from the Hong Kong Special Administrative Region.

References

1. Arif, Z.H., Mahmoud, M.A., Abdulkareem, K.H., Mohammed, M.A., Al-Mhiqani, M.N., Mutlag, A.A., Damaševičius, R.: Comprehensive review of machine learning (ml) in image defogging: Taxonomy of concepts, scenes, feature extraction, and classification techniques. IET Image Processing **16**(2), 289–310 (2022) 2
2. Ayzenberg, L., Giryes, R., Greenspan, H.: Dinov2 based self supervised learning for few shot medical image segmentation. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024) 4

3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 4, 14
4. Brophy, T., Mullins, D., Parsi, A., Horgan, J., Ward, E., Denny, P., Eising, C., Deegan, B., Glavin, M., Jones, E.: A review of the impact of rain on camera-based perception in automated driving systems. IEEE Access **11**, 67040–67057 (2023) 2
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021) 4
6. Chen, C., Miao, J., Wu, D., Zhong, A., Yan, Z., Kim, S., Hu, J., Liu, Z., Sun, L., Li, X., et al.: Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. Medical Image Analysis **98**, 103310 (2024) 4
7. Ch'Ng, C.K., Chan, C.S.: Total-text: A comprehensive dataset for scene text detection and recognition. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR). vol. 1, pp. 935–942. IEEE (2017) 2, 10
8. Guan, T., Gu, C., Lu, C., Tu, J., Feng, Q., Wu, K., Guan, X.: Industrial scene text detection with refined feature-attentive network. IEEE Transactions on Circuits and Systems for Video Technology **32**(9), 6073–6085 (2022) 1
9. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2315–2324 (2016) 2, 4, 11
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022) 5
11. Jiang, J., Zuo, Z., Wu, G., Jiang, K., Liu, X.: A survey on all-in-one image restoration: Taxonomy, evaluation and future trends. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 2
12. Jose, C., Moutakanni, T., Kang, D., Baldassarre, F., Darcet, T., Xu, H., Li, D., Szafraniec, M., Ramamonjisoa, M., Oquab, M., et al.: Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24905–24916 (2025) 4
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023) 4
14. Li, C., Liu, W., Guo, R., Yin, X., Jiang, K., Du, Y., Du, Y., Zhu, L., Lai, B., Hu, X., et al.: Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. arXiv preprint arXiv:2206.03001 (2022) 4, 14
15. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3041–3050 (2023) 4
16. Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X.: Real-time scene text detection with differentiable binarization. In: Proceedings of the AAAI conference on artificial intelligence. pp. 11474–11481 (2020) 2, 3, 10, 11, 12, 14
17. Liu, S.Y., Wang, C.Y., Yin, H., Molchanov, P., Wang, Y.C.F., Cheng, K.T., Chen, M.H.: Dora: Weight-decomposed low-rank adaptation. In: Forty-first International Conference on Machine Learning (2024) 5
18. Long, S., He, X., Yao, C.: Scene text detection and recognition: The deep learning era. International Journal of Computer Vision **129**(1), 161–184 (2021) 2

19. Luo, S., Chen, W., Tian, W., Liu, R., Hou, L., Zhang, X., Shen, H., Wu, R., Geng, S., Zhou, Y., et al.: Delving into multi-modal multi-task foundation models for road scene understanding: From learning paradigm perspectives. IEEE Transactions on Intelligent Vehicles (2024) 2
20. Luo, Y., Feinglass, J., Gokhale, T., Lee, K.C., Baral, C., Yang, Y.: Grounding stylistic domain generalization with quantitative domain shift measures and synthetic scene images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7303–7313 (2024) 2
21. Ma, X., Wu, Q., Zhao, X., Zhang, X., Pun, M.O., Huang, B.: Sam-assisted remote sensing imagery semantic segmentation with object and boundary constraints. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–16 (2024) 4
22. Mao, Y., Ge, Y., Fan, Y., Xu, W., Mi, Y., Hu, Z., Gao, Y.: A survey on lora of large language models. Frontiers of Computer Science **19**(7), 197605 (2025) 5
23. Miao, J., Chen, C., Yuan, Y., Li, Q., Heng, P.A.: Sam-driven cross prompting with adaptive sampling consistency for semi-supervised medical image segmentation. Medical Image Analysis p. 103973 (2026) 4
24. Naiemi, F., Ghods, V., Khalesi, H.: Scene text detection and recognition: a survey. Multimedia Tools and Applications **81**(14), 20255–20290 (2022) 1
25. Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., Luo, Z., Pal, U., Rigaud, C., Chazalon, J., et al.: Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR). vol. 1, pp. 1454–1459. IEEE (2017) 10
26. Neat, L., Peng, R., Qin, S., Manduchi, R.: Scene text access: A comparison of mobile ocr modalities for blind users. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. pp. 197–207 (2019) 1
27. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 4
28. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024) 4
29. Shao, Z., Su, Y., Zhou, Y., Meng, F., Zhu, H., Liu, B., Yao, R.: Ct-net: Arbitrary-shaped text detection via contour transformer. IEEE Transactions on Circuits and Systems for Video Technology **34**(3), 1815–1826 (2023) 10, 11
30. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025) 4
31. Song, Y., Sun, P., Liu, H., Li, Z., Song, W., Xiao, Y., Zhou, X.: Scene-driven multimodal knowledge graph construction for embodied ai. IEEE Transactions on Knowledge and Data Engineering **36**(11), 6962–6976 (2024) 1
32. Su, H., Li, Y., Xu, Y., Fu, X., Liu, S.: A review of deep-learning-based super-resolution: From methods to applications. Pattern Recognition **157**, 110935 (2025) 2
33. Su, Y., Chen, Z., Du, Y., Ji, Z., Hu, K., Bai, J., Gao, X.: Explicit relational reasoning network for scene text detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 7069–7077 (2025) 2, 4, 10, 11
34. Su, Y., Chen, Z., Shao, Z., Du, Y., Ji, Z., Bai, J., Zhou, Y., Jiang, Y.G.: Lranet: Towards accurate and efficient scene text detection with low-rank approximation network. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 4979–4987 (2024) 2, 4, 10, 11, 12

35. Su, Y., Shao, Z., Zhou, Y., Meng, F., Zhu, H., Liu, B., Yao, R.: Textdct: Arbitrary-shaped text detection via discrete cosine transform mask. IEEE Transactions on Multimedia **25**, 5030–5042 (2022) 10, 11
36. Sun, H., Cai, C., Zhuang, H., Lee, K.A., Chau, L.P., Wang, Y.: Edvd-llama: Explainable deepfake video detection via multimodal large language model reasoning. arXiv preprint arXiv:2510.16442 (2025) 14
37. Vaidya, S., Sharma, A.K., Gatti, P., Mishra, A.: Show me the world in my language: Establishing the first baseline for scene-text to scene-text translation. In: International Conference on Pattern Recognition. pp. 312–328. Springer (2024) 1
38. Wang, R., Chen, H., Zhu, Y., Xu, J., Cao, X., Zhu, Z., Qian, S., Gao, C., Liu, L., Sang, N.: S3inet: Semantic-information space sharing interaction network for arbitrary shape text detection. IEEE Transactions on Neural Networks and Learning Systems (2025) 2, 3, 10, 11, 12, 14
39. Wang, Y., Bian, Z.P., Hou, J., Chau, L.P.: Convolutional neural networks with dynamic regularization. IEEE Transactions on Neural Networks and Learning Systems **32**(5), 2299–2304 (2020) 2
40. Wang, Y., Bian, Z.P., Zhou, Y., Chau, L.P.: Rethinking and designing a high-performing automatic license plate recognition approach. IEEE transactions on intelligent transportation systems **23**(7), 8868–8880 (2021) 2
41. Wang, Y., Hou, J., Hou, X., Chau, L.P.: A self-training approach for point-supervised object detection and counting in crowds. IEEE Transactions on Image Processing **30**, 2876–2887 (2021) 2
42. Wei, H., Sun, Y., Li, Y.: Deepseek-ocr: Contexts optical compression. arXiv preprint arXiv:2510.18234 (2025) 4, 14
43. Xiong, Y., Zhou, C., Xiang, X., Wu, L., Zhu, C., Liu, Z., Suri, S., Varadarajan, B., Akula, R., Iandola, F., et al.: Efficient track anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11513–11524 (2025) 4
44. Xue, C., Zhang, W., Hao, Y., Lu, S., Torr, P.H., Bai, S.: Language matters: A weakly supervised vision-language pre-training approach for scene text detection and spotting. In: European Conference on Computer Vision. pp. 284–302. Springer (2022) 2
45. Yan, Z., Li, J., Li, X., Zhou, R., Zhang, W., Feng, Y., Diao, W., Fu, K., Sun, X.: Ringmo-sam: A foundation model for segment anything in multimodal remote-sensing images. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–16 (2023) 4
46. Yang, S., Xie, L., Ran, X., Lei, J., Qian, X.: Pragmatic degradation learning for scene text image super-resolution with data-training strategy. Knowledge-Based Systems **285**, 111349 (2024) 2
47. Yuliang, L., Lianwen, J., Shuaitao, Z., Sheng, Z.: Detecting curve text in the wild: New dataset and new solution. arXiv preprint arXiv:1712.02170 (2017) 2, 9
48. Zhang, Q., Chen, M., Bukharin, A., Karampatziakis, N., He, P., Cheng, Y., Chen, W., Zhao, T.: Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. arXiv preprint arXiv:2303.10512 (2023) 5
49. Zhang, S.X., Zhu, X., Chen, L., Hou, J.B., Yin, X.C.: Arbitrary shape text detection via segmentation with probability maps. IEEE transactions on pattern analysis and machine intelligence **45**(3), 2736–2750 (2022) 2, 3, 10, 11, 12, 14
50. Zhang, S.X., Zhu, X., Yang, C., Wang, H., Yin, X.C.: Adaptive boundary proposal network for arbitrary shape text detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1305–1314 (2021) 2, 4, 10, 11, 12

51. Zhao, Q., Li, G., He, B., Shen, R.: Deep learning for low-light vision: A comprehensive survey. IEEE Transactions on Neural Networks and Learning Systems (2025) 2
52. Zhao, X., Feng, W., Zhang, Z., Lv, J., Zhu, X., Lin, Z., Hu, J., Shao, J.: Cbnet: A plug-and-play network for segmentation-based scene text detection. International Journal of Computer Vision **132**(8), 3119–3138 (2024) 10, 11, 12