ESTextSpotter: Towards Better Scene Text Spotting with Explicit Synergy in Transformer

Mingxin Huang[1†] Jiaxin Zhang[2†] Dezhi Peng[1] Hao Lu[3] Can Huang[2]
Yuliang Liu[3] Xiang Bai[3] Lianwen Jin[1*]
[1]South China University of Technology [2]ByteDance
[3]Huazhong University of Science and Technology
eelwjin@scut.edu.cn

Abstract

In recent years, end-to-end scene text spotting approaches are evolving to the Transformer-based framework. While previous studies have shown the crucial importance of the intrinsic synergy between text detection and recognition, recent advances in Transformer-based methods usually adopt an implicit synergy strategy with shared query, which can not fully realize the potential of these two interactive tasks. In this paper, we argue that the explicit synergy considering distinct characteristics of text detection and recognition can significantly improve the performance text spotting. To this end, we introduce a new model named Explicit Synergy-based Text Spotting Transformer framework (ESTextSpotter), which achieves explicit synergy by modeling discriminative and interactive features for text detection and recognition within a single decoder. Specifically, we decompose the conventional shared query into task-aware queries for text polygon and content, respectively. Through the decoder with the proposed vision-language communication module, the queries interact with each other in an explicit manner while preserving discriminative patterns of text detection and recognition, thus improving performance significantly. Additionally, we propose a task-aware query initialization scheme to ensure stable training. Experimental results demonstrate that our model significantly outperforms previous state-of-the-art methods. Code is available at `https://github.com/mxin262/ESTextSpotter`.

1. Introduction

End-to-end text spotting, aiming at building a unified framework for text detection and recognition in natural scenes, has received great attention in recent years [32, 25, 33]. Intuitively, the position and shape of the text in the detection can help the text recognition accurately extract the content of the text. Similarly, the position and classification information in recognition can also guide the detector to distinguish between different text instances and the background. Such mutual interaction and cooperation between text detection and recognition are recently known as scene text spotting synergy [17], which aims to produce a combined effect greater than the sum of their separate effects. Indeed, synergy is the key to the success in literature.

In the past few years, many methods attempt to join text detection and recognition by proposing a new Region-of-Interest (RoI) operation to achieve the synergy between text detection and text recognition [32, 12, 50, 33, 52], as shown

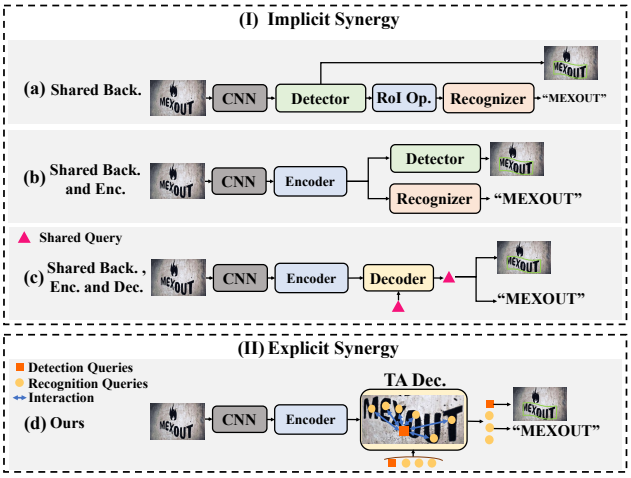

Figure 1: Comparison of implicit and explicit synergy between text detection and recognition. Implicit synergy is achieved by sharing parameters and features. Explicit synergy is attained by explicitly modeling discriminative and interactive features. Back.: backbone. Enc. (Dec.): encoder (decoder). TA Dec.: task-aware decoder.

[†]Equal contribution.
[*]Corresponding author.

in Figure 1(a). They follow the classical two-stage pipeline, which first locates the text instance and then extracts the text content in the corresponding region of interest (RoI). However, the interaction between detection and recognition is insufficient through sharing a backbone, as observed in recent research [17]. A recent study, TESTR [68], develops dual-decoder framework to further share an encoder, but there is still a lack of interaction between the two tasks, as presented in Figure 1(b). Therefore, some researchers [21, 63] begin to explore better synergy based on the Transformer [49] architecture. For instance, TTS [21] takes a step toward unifying the detector and recognizer into a single decoder with shared query for both two tasks as illustrated in Figure 1(c). DeepSolo [63] further adopts a group of shared queries to encode the characteristics of text. Although these approaches [21, 63] develop a more concise and unified framework, they fail to consider distinct feature patterns of these two tasks. We formulate the above-mentioned methods as utilizing an implicit synergy that shares parameters and features between the detection and recognition, but lacks explicit modeling between them, as shown in Figure 1(I). The full potential of two tasks can not be realized by implicit synergy alone without considering the unique characteristics of each task [47, 60]. For instance, while DeepSolo has demonstrated promising end-to-end results on Total-Text [7], its detection performance falls short of that achieved by the dedicated detection method [48].

In this paper, we propose an Explicit synergy Text Spotting Transformer framework, termed ESTextSpotter, stepping toward explicit synergy between text detection and recognition. Compared to previous implicit synergy, ESTextSpotter explicitly models discriminative and interactive features for text detection and recognition within a single decoder, as illustrated in Figure 1(d). Technically, we design a set of task-aware queries to model the different feature patterns of text detection and recognition, which include detection queries encoding the position and shape information of the text instance, and recognition queries encoding the position and semantics information of the character. The position information of the character is obtained through an attention mechanism similar to previous works [11, 57]. Then, detection queries and recognition queries are sent into a task-aware decoder that is equipped with a vision-language communication module to enhance explicit synergy. Previous works [68, 21, 63] have used learnable embeddings to initialize the queries. However, these randomly initialized parameters will disrupt the training of the vision-language communication module. Therefore, we propose a task-aware queries initialization (TAQI) to promote stable training of the vision-language communication module. Besides, inspired by [23, 64], we also employ a denoising training strategy to expedite convergence.

Extensive experiments demonstrate the effectiveness of our method: 1) For text detection, ESTextSpotter significantly outperforms previous detection methods by an average of 3.0% in terms of the H-mean on two arbitrarily-shaped text datasets, 1.8% on two multi-oriented datasets, and 3.0% on Chinese and multi-lingual datasets; 2) For English text spotting, ESTextSpotter consistently outperforms previous methods by large margins; 3) ESTextSpotter also significantly outperforms previous methods on multilingual text spotting including Chinese text (ReCTS), African Amharic text (HUST-Art), and Vietnamese text (VinText), with an average of 4.8% in terms of the end-to-end H-mean.

In conclusion, our contributions can be summarized as follows.

- We introduce ESTextSpotter, a simple yet efficient Transformer-based approach for text spotting that adopts task-aware queries within a single decoder, which allows it to effectively realize explicit synergy of text detection and recognition, thereby unleashing the potential of these two tasks.
- We propose a vision-language communication module designed to enhance explicit synergy, which utilizes a novel collaborative cross-modal interaction between text detection and recognition. Moreover, we introduce a task-aware query initialization module to guarantee stable training of the module.
- We achieve significant improvements over state-of-the-art methods across eight challenging scene text spotting benchmarks.

2. Related Work

End-to-End Scene Text Spotting. Classical methods [51, 18, 27] have some limitations in addressing scene text spotting, such as error accumulation, sub-optimization, and low inference efficiency. To overcome these problems, a paradigm shift has been witnessed from shallow learning to end-to-end learning. In particular, Li *et al*. [24] integrated detection and recognition into a unified end-to-end framework. However, this method mainly handles horizontal texts. Some researchers introduced special RoI operations, such as Text-Align [15] and RoI-Rotate [32], to sample the oriented text features into regular ones for text recognition. Liao *et al*. [37] proposed Mask TextSpotter, which introduces a character segmentation module to take the advantage of character-level annotations, to solve the problem of arbitrarily-shaped scene text spotting. TextDragon [12] further proposed RoISlide to fuse features from the predicted segments for text recognition. Qin *et al*. [44] proposed RoI Masking to suppress the background noise by multiplying segmentation masks with features. Wang *et al*. [50] used a boundary detector to cooperate with the use of Thin-Plate-Spline (TPS) [4]. Mask TextSpotter v3 [26] proposed a

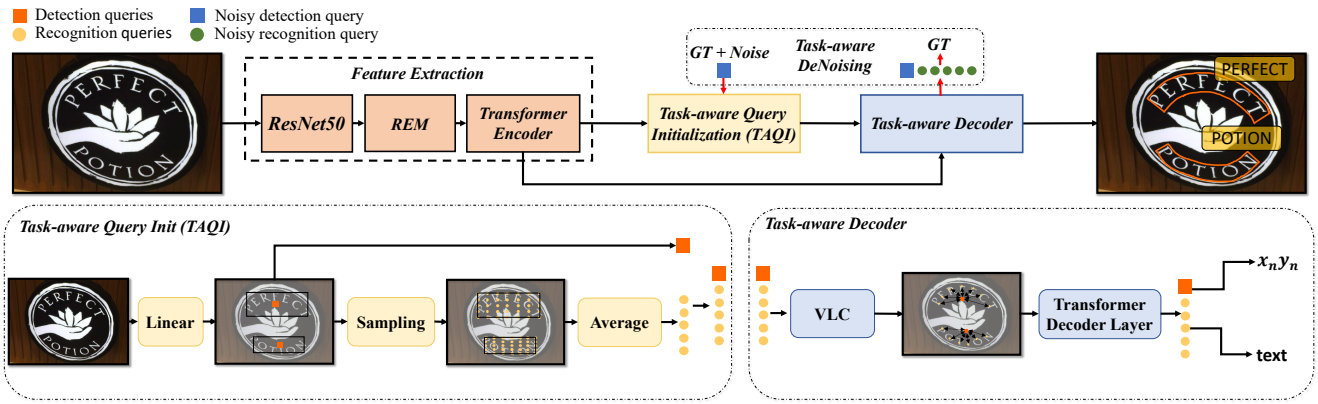

Figure 2: The framework of the proposed ESTextSpotter. The image features are extracted in the feature extraction process. Then the Task-aware Query Initialization is used to generate the task-aware queries including detection and recognition queries. Then the task-aware queries are sent into the task-aware decoder to obtain the detection and recognition results simultaneously. REM is the Receptive Enhancement Module. VLC means the vision-language communication module. The red arrow means only used in the training stage.

Segmentation Proposal Network (SPN) to generate accurate proposals for arbitrarily-shaped text. MANGO [43] developed a Mask Attention module to coarsely localize characters, which requires character-level annotations. ABCNet [33] and its improved version ABCNet v2 [35] used the parametric bezier curve to model the curved text and developed Bezier-Align for rectifying curved text. The methods discussed above mainly focus on designing shape-aware Region of Interest (RoI) sampling, while merely achieving synergy by sharing the backbone.

Text Spotting Transformer. To further enhance the interaction between detection and recognition, TETSR [68] developed a dual-decoder framework to share both backbone and encoder between two tasks, and only detection and recognition head are isolated. SwinTextSpotter [17] further proposed a Recognition Conversion to implicitly guide the recognition head through incorporating the detection and back-propagate recognition information to the detector. TTS [21] attempted to unify the detector and recognizer in a single decoder using a shared query. To encode the characteristics of text in the queries, DeepSolo [63] utilized a group of point queries based on the center line. Similarly, SPTS [42] adopted an auto-regressive framework [6] that most parameters are shared between text detection and recognition.

Although the Transformer has shown great potential in text spotting, current methods still have limitations. Firstly, the dual-decoder framework [68] lacks interaction between text detection and recognition, which limits the performance. Secondly, the shared query in the single decoder framework [21, 63] does not fully consider the distinct feature patterns of these two tasks. Note that, while a closely related work, SwinTextSpotter [17], also attempts to explore the synergy between text detection and recognition, it does not fully achieve explicit synergy. This is because it back-propagates recognition information to the detector without explicitly modeling the relationship between text detection and recognition.

3. Methodology

In this paper, we propose an Explicit synergy-based Text Spotting Transformer framework, termed ESTextSpotter. The key idea of ESTextSpotter is to explicitly model discriminative and interactive features for text detection and recognition within a single decoder. The overall architecture is shown in Figure 2. After obtaining image features through the feature extraction process consisting of ResNet50, receptive enhancement module (REM), and Transformer encoder, task-aware queries are generated using the Task-aware Query Initialization module (TAQI), which includes detection and recognition queries. To achieve better explicit synergy, these queries are then sent into the task-aware decoder to explicitly model discriminative and interactive features for text detection and recognition simultaneously. During training, inspired by previous works [23, 64], we utilize a task-aware DeNoising training strategy to accelerate convergence. Detailed implementations will be provided in the following subsections.

3.1. Receptive Enhancement Module

Following previous works [25, 35, 68], we adopt ResNet50 [13] as our backbone. To enhance the receptive field of the features, we send the feature map res_5 output from the ResNet50 to the receptive enhancement module (REM), which uses a convolutional layer to downsample the feature map. Then, we send the output of the REM, as well as the feature maps res_3 to res_5, to the Transformer encoder [69] to model long-range dependencies across var-

ious scales. Finally, the output of the Transformer encoder is fed into the subsequent modules.

3.2. Task-aware Query Initialization

Previous works [68, 21, 62, 63] have utilized learnable embeddings to initialize the queries. However, these randomly initialized parameters will disrupt the training of the vision-language communication module. Therefore, we propose task-aware query initialization (TAQI) to improve the stability of the vision-language communication module during training. Firstly, we use a linear layer to generate text classification scores from the output of the Transformer encoder. Then, we select the top N features based on the text classification scores, and these features are sent into a linear layer to initialize proposals for detection queries. For recognition queries, we sample the features $\mathbf{F_p} \in \mathbb{R}^{N\times H\times T\times C}$ from the proposals and average over the height dimension to initialize recognition queries. Here, N represents the maximum number of predictions identified in DETR [5], while T represents the length of recognition queries, and C represents the feature dimension. Benefiting from the decomposition of the conventional shared query, TAQI encodes the boundary and content information into the detection and recognition queries, respectively.

3.3. Task-aware Decoder

After obtaining the task-aware queries $\mathbf{S} \in \mathbb{R}^{N\times(T+1)\times C}$, including detection queries $\mathbf{G} \in \mathbb{R}^{N\times C}$ and recognition queries $\mathbf{R} \in \mathbb{R}^{N\times T\times C}$, they are sent to the task-aware decoder to interact and mutually promote each other. We first enhance the explicit synergy between task-aware queries from a cross-modal perspective in the proposed vision-language communication module as illustrated in Figure 3. A Language Conversion is designed to extract semantic features in recognition queries and map them into language vectors $\mathbf{L} \in \mathbb{L}^{N\times(T+1)\times C}$, which is defined as follows:

$$\mathbf{P} = \texttt{softmax}(\mathbf{W_1}\mathbf{R}), \tag{1}$$

$$\mathbf{L} = \texttt{cat}(\mathbf{G}, \mathbf{W_2}\mathbf{P}), \tag{2}$$

where $\mathbf{W_1} \in \mathbb{R}^{C\times U}$ and $\mathbf{W_2} \in \mathbb{R}^{U\times C}$ are trainable weights. U indicates the character class number. $\texttt{cat}$ is the concatenation operation. Then the task-aware queries $\mathbf{S}$ and language vectors $\mathbf{L}$ are sent to a vision-language attention module, which is formalized as:

$$\mathbf{M}_{ij} = \begin{cases} 0, & i \neq j, \\ -\infty, & i = j. \end{cases} \tag{3}$$

$$\mathbf{F} = \texttt{softmax}(\frac{(\mathbf{S}+\text{PE}(\mathbf{S}))(\mathbf{L}+\text{PE}(\mathbf{L}))^T}{\sqrt{D}} + \mathbf{M})\mathbf{L}. \tag{4}$$

PE indicates the position encoding used by DETR [5]. The attention mask $\mathbf{M}$ is designed to prevent the queries from over-focusing "itself".

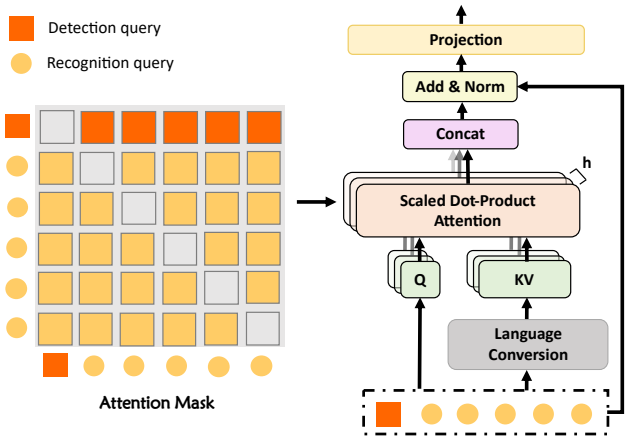

Figure 3: **Illustration of the Vision-language communication module**. The language conversion extracts the semantic features in recognition queries. Then the visual and semantic information can interact in a cross-modal perspective. h is the number of parallel attention heads.

After exchanging vision-language information, we send the task-aware queries to the Transformer decoder layer. Consistent with prior research works [68, 62, 63], we first initially incorporate an intra-group self-attention module to enhance the relationship between task-aware queries within one text instance. Subsequently, we use an inter-group self-attention module to model the relationship between distinct text instances. The outputs of these modules are then fed into a multi-scale deformable cross-attention module to extract the text features from the Transformer encoder. Finally, we employ two linear layers to predict the detection and recognition results.

During the decoding stage, the detection queries extract the position and shape information of text instances, while the recognition queries comprise the semantic and positional information of characters. When explicitly modeling the relationship in the task-aware decoder, the positional information of the character available in the recognition queries can assist the detection queries in accurately locating the text. Similarly, the positional and shape information of the text instance present in the detection queries can help the recognition queries in extracting character features. As a result, this explicit synergy between detection and recognition queries unleashes the potential of both text detection and recognition while also preserving the distinct feature patterns in the detection and recognition queries.

Detection and Recognition Format. In contrast to previous works that utilize serial point queries to model the curve [68, 62, 63] or freeze the model weights to train a segmentation head [21], we develop a simpler approach to generate the detection results. We send the detection queries

into two feed-forward layers, in which one predicts the proposals (x, y, h, w), while the other predicts Z polygon offsets $(\Delta x_1, \Delta y_1, \Delta x_2, \Delta y_2, ..., \Delta x_Z, \Delta y_Z)$ based on the center point (x, y) of each proposal. Z is 16. Reconstructing the polygon can be formulated as:

$$x_i = x + \Delta x_i \,, \tag{5}$$

$$y_i = y + \Delta y_i \,. \tag{6}$$

In this way, we can predict the detection result directly through detection queries without serial control points or freezing the model weights to train a segmentation head. Following the decoding process, the recognition queries can efficiently extract character features. We utilize a linear layer to convert the recognition queries into characters, similar to [68].

3.4. Optimization

Task-aware DeNoising. Recently, some researchers [23, 64] propose the DeNoising training to accelerate the convergence of the DETR [5]. However, these methods are specifically designed for detection. Therefore, we develop a Task-aware DeNoising (TADN) strategy for text spotting to accelerate the convergence, as presented in Figure 2. Following previous works [23, 64], we add center shifting and box scaling in ground truth boxes, termed noise boxes. The noise boxes are transformed into noise detection queries by linear layers, and the noise recognition queries are initialized by TAQI. The noise detection and recognition queries are concatenated and sent to the task-aware decoder, which is responsible for reconstructing the ground truth boxes and obtaining the corresponding recognition results. TADN more focuses on text spotting rather than detection, as opposed to previous denoising training methods [23, 64].

Loss. The training process of ESTextSpotter is a set prediction problem that uses a fixed number of outputs to match the ground truths. Inspired by the DETR-like methods [5, 69, 31], we utilize the Hungarian algorithm [22] to perform pairwise matching and minimize the prediction-ground truth matching cost $\mathcal{C}_{match}$ as:

$$\hat{\sigma} = \arg\min_{\sigma} \sum_{i=1}^{N} \mathcal{C}_{match}(Y_i, \hat{Y}_{\sigma(i)}) \,, \tag{7}$$

where Y_i is the ground truth and $\hat{Y}_{\sigma(i)}$ is the prediction. N is the number of the predictions indexed by $\sigma(i)$. The cost function $\mathcal{C}_{match}$ is defined as:

$$\mathcal{C}_{match}(Y_i, \hat{Y}_{\sigma(i)}) = \lambda_c \mathcal{C}_c(\hat{p}_{\sigma(i)}(c_i)) + \mathbb{1}_{\{c_i \neq \emptyset\}} \lambda_b \mathcal{C}_b(b_i, \hat{b}_{\sigma(i)}) \,, \tag{8}$$

where c_i and b_i are the ground truth class and bounding box, and $\hat{b}_{\sigma(i)}$ represents the prediction of bounding box. $\hat{p}_{\sigma(i)}(c_i)$ is the probability of prediction for class c_i. λ_c and λ_b are the weights for the classification and bounding box. After the Hungarian algorithm, the prediction and ground truth can be one-to-one matched. The training loss is as follows:

$$\begin{aligned} \mathcal{L}(Y_i, \hat{Y}_{\sigma(i)}) = \alpha_c \mathcal{L}_c(\hat{p}_{\sigma(i)}(c_i)) + \mathbb{1}_{\{c_i \neq \emptyset\}} \mathcal{L}_b(b_i, \hat{b}_{\sigma(i)}) \\ + \mathbb{1}_{\{c_i \neq \emptyset\}} \alpha_r \mathcal{L}_r(r_i, \hat{r}_{\sigma(i)}) + \mathbb{1}_{\{c_i \neq \emptyset\}} \alpha_p \mathcal{L}_p(p_i, \hat{p}_{\sigma(i)}) \,, \end{aligned} \tag{9}$$

where α_c, α_b, α_p, and α_r are the loss weights for the classification, bounding box, polygon, and recognition, respectively. The classification loss $\mathcal{L}_c$ is the focal loss [30]. The bounding box loss $\mathcal{L}_b$ consists of the ℓ_1 loss and the GIoU loss [45]. The polygon loss uses the ℓ_1 loss as well. The recognition loss is the standard cross entropy loss.

4. Experiments and Results

We conduct experiments on common benchmarks to evaluate ESTextSpotter, including multi-oriented ICDAR2015 [19] and MSRA-TD500 [61], multilingual datasets ReCTS [65], Vintext [41], HUST-ART [9], and ICDAR2019-MLT [39], arbitrarily shaped datasets Total-Text [7], and SCUT-CTW1500 [34].

4.1. Implementation Details

We pre-train the model on a combination of Curved SynthText [33], ICDAR-MLT [40], and the corresponding datasets with 240K iterations. The base learning rate is 1×10^{-4} and reduced to 1×10^{-5} at the 180K-th iteration and 1×10^{-6} at 210K-th iteration. Then, the model is pre-trained on the Total-Text [7], ICDAR 2013 [20], and ICDAR-MLT and fine-tuned on the corresponding real datasets. For Chinese and Vietnamese datasets, we follow the training strategies in previous works [41, 35] to train the model. We use $N = 100$ as the maximum number of predictions. The max length of recognition queries T is 25. The weight for the classification loss α_c is 2.0. The weight of the ℓ_1 loss is 5.0 and of the GIoU loss is 2.0. The polygon loss weight α_p and the recognition loss weight α_r are both set to 1.0. The focal loss parameters α and γ are 0.25 and 2.0, respectively. The number of both encoder and decoder layers is 6. The inference speed is tested on a single NVIDIA GeForce RTX 3090.

The data augmentation strategies used are also kept the same as previous works [68, 33, 35] as follows: 1) random resizing with the shorter size chosen from 640 to 896 pixels (with an interval of 32), and the longest size is constrained within 1600 pixels; 2) random cropping, which ensures that text is not being cut; 3) random rotation, which rotates the images with an angle in range of $[-45°, 45°]$. For testing, we resize the shorter size of the image to 1000 pixels while keeping the longest size of the image within 1824 pixels.

Table 1: Detection results on the Total-Text, SCUT-CTW1500, MSRA-TD500, and ICDAR 2015 datasets. AAvg. means the average in arbitrarily-shaped text. MAvg. means the average in multi-oriented text. Bold indicates SOTA, and underline indicates the second best.

Methods	Total-Text			SCUT-CTW1500			MSRA-TD500			ICDAR 2015			ReCTS			AAvg.	MAvg.
	R	P	H	R	P	H	R	P	H	R	P	H	R	P	H		
TextDragon [12]	75.7	85.6	80.3	82.8	84.5	83.6	–	–	–	–	–	–	–	–	–	82.0	–
PSENet–1s [53]	78.0	84.0	80.9	79.7	84.8	82.2	–	–	–	85.5	88.7	87.1	83.9	87.3	85.6	81.6	–
CRAFT [1]	79.9	87.6	83.6	81.1	86.0	83.5	78.2	88.2	82.9	84.3	89.8	86.9	–	–	–	83.6	84.9
PAN [55]	81.0	89.3	85.0	81.2	86.4	83.7	83.8	84.4	84.1	81.9	84.0	82.9	–	–	–	84.4	83.5
DBNet [28]	82.5	87.1	84.7	80.2	86.9	83.4	77.7	76.6	81.9	82.7	88.2	85.4	–	–	–	84.1	83.7
DRRG [66]	84.9	86.5	85.7	83.0	85.9	84.5	82.3	88.1	85.1	84.7	88.5	86.6	–	–	–	85.1	85.9
CounterNet [58]	83.9	86.9	85.4	84.1	83.7	83.9	–	–	–	86.1	87.6	86.9	–	–	–	84.7	–
FCENet [70]	82.5	89.3	85.8	83.4	87.6	85.5	–	–	–	82.6	90.1	86.2	–	–	–	85.7	–
PCR [8]	82.0	88.5	85.2	82.3	87.2	84.7	83.5	90.8	87.0	–	–	–	–	–	–	85.0	–
MOST [14]	–	–	–	–	–	–	82.7	90.4	86.4	87.3	89.1	88.2	–	–	–	–	87.3
TextBPN[67]	85.2	90.7	87.9	83.6	86.5	85.0	84.5	86.6	85.6	–	–	–	–	–	–	86.5	–
ABCNet v2[35]	84.1	90.2	87.0	83.8	85.6	84.7	81.3	89.4	85.2	86.0	90.4	88.1	87.5	93.6	90.4	85.9	86.7
PAN++[35]	81.0	89.9	85.3	81.1	87.1	84.0	85.6	91.4	88.4	83.9	91.4	87.5	–	–	–	84.7	88.0
DBNet++[29]	83.2	88.9	86.0	82.8	87.9	85.3	83.3	91.5	87.2	83.9	90.9	87.3	–	–	–	85.7	87.3
FSGNet[48]	85.7	90.7	88.1	82.4	88.1	85.2	84.8	91.6	88.1	86.7	91.1	88.8	–	–	–	86.7	88.5
TESTR[68]	81.4	**93.4**	86.9	82.6	**92.0**	87.1	–	–	–	**89.7**	90.3	90.0	–	–	–	87.0	–
DeepSolo[63]	82.1	93.1	87.3	–	–	–	–	–	–	87.4	92.8	90.0	–	–	–	–	–
ESTextSpotter (Ours)	**88.1**	92.0	**90.0**	**88.6**	91.5	**90.0**	**86.3**	**92.9**	**89.5**	89.6	**92.5**	**91.0**	**91.3**	**94.1**	**92.7**	**90.0**	**90.3**

Table 2: Detection results on MLT19 and language-wise performance. CRAFTS (paper) means that the result comes from the paper [2]. The result of CRAFTS* comes from the official ICDAR19-MLT website.

Method	R	P	H	AP	Arabic	Latin	Chinese	Japanese	Korean	Bangla	Hindi
PSENet [53]	59.59	73.52	65.83	52.73	43.96	65.77	38.47	34.47	51.73	34.04	47.19
RRPN [38]	62.95	77.71	69.56	58.07	35.88	68.01	33.31	36.11	45.06	28.78	40.00
CRAFTS* [2]	62.73	81.42	70.86	56.63	43.97	72.49	37.20	42.10	54.05	38.50	**53.50**
CRAFTS (paper) [2]	70.1	81.7	75.5	–	–	–	–	–	–	–	–
Single-head TextSpotter [26]	61.76	83.75	71.10	58.76	51.12	73.56	40.41	41.22	56.54	39.68	49.00
Multiplexed TextSpotter [16]	63.16	**85.53**	72.66	60.46	51.75	73.55	43.86	42.43	57.15	**40.27**	51.95
DBNet [28]	64.0	78.3	70.4	–	–	–	–	–	–	–	–
DBNet++ [29]	65.4	78.6	71.4	–	–	–	–	–	–	–	–
ESTextSpotter (Ours)	**75.5**	83.37	**79.24**	**72.52**	**52.00**	**77.34**	**48.20**	**48.42**	**63.56**	38.26	50.83

4.2. Comparison with State-of-the-Arts

Multi-oriented Text. We conduct experiments on ICDAR2015 and MSRD-TD500 [61] to evaluate the robustness of our method for multi-oriented text. The detection results are presented in Table 1. Our method achieves the highest H-mean score of 91.0% on the ICDAR2015 dataset, outperforming DeepSolo by 1.0%. On the MSRA-TD500 dataset, ESTextSpotter achieves an accuracy of 89.5%. These results demonstrate the robustness of our method for detecting long, straight text. The end-to-end recognition results on the ICDAR2015 are shown in Table 3. Our method outperforms previous methods on all lexicon settings. Notably, in the strong lexicon setting, ESTextSpotter achieves 87.5% in terms of the Hmean, 2.3% higher than the TESTR and TTS. In weak and generic lexicon, ESTextSpotter outperforms the state-of-the-art implicit synergy method DeepSolo by 1.1% and 1.2%, respectively. It demonstrates the effectiveness of the proposed explicit synergy.

Arbitrarily-Shaped Text. We test our method on two arbitrarily-shaped text benchmarks (Total-Text and CTW1500) to verify the generalization ability of our approach for arbitrarily-shaped scene text spotting. For text detection task, as shown in Table 1, our method outperforms the previous state-of-the-art model with 90.0% in terms of the H-mean metric on Total-Text dataset, 2.6% higher than the DeepSolo. On the CTW1500 dataset, our method also achieves 90.0%, which also significantly outperforms previous methods. The end-to-end scene text spotting results are shown in Table 3, ESTextSpotter significantly surpasses the TTS by a large margin (2.6% without lexicon and 0.8% on "Full" lexicon) on TotalText. On the CTW1500 dataset, ESTextSpotter outperforms all previous best models by 0.7% without lexicon and 2.4% on 'Full' lexicon. From Tables 1 to 3, it can be seen that our method consistently achieves the best results for text detection and text spotting.

Table 3: End-to-end text spotting results on Total-Text, SCUT-CTW1500, ICDAR2015 and ReCTS. 'None' means lexicon-free. 'Full' indicates that we use all the words that appeared in the test set. 'S', 'W', and 'G' represent recognition with 'Strong', 'Weak', and 'Generic' lexicons, respectively.

Methods	Total-Text		SCUT-CTW1500		ICDAR 2015 End-to-End			ReCTS	FPS
	None	Full	None	Full	S	W	G	1-NED	
Mask TextSpotter [25]	65.3	77.4	–	–	83.0	77.7	73.5	67.8	–
FOTS [32]	–	–	21.1	39.7	83.6	79.1	65.3	50.8	–
TextDragon [12]	48.8	74.8	39.7	72.4	82.5	78.3	65.2	–	–
Text Perceptron [12]	69.7	78.3	57.0	–	80.5	76.6	65.1	–	–
ABCNet [33]	64.2	75.7	45.2	74.1	–	–	–	–	17.9
Mask TextSpotter v3 [26]	71.2	78.4	–	–	83.3	78.1	74.2	–	–
PGNet [52]	63.1	–	–	–	83.3	78.3	63.5	–	35.5
MANGO [43]	72.9	83.6	58.9	78.7	81.8	78.9	67.3	–	4.3
ABCNet v2 [35]	70.4	78.1	57.5	77.2	82.7	78.5	73.0	62.7	10.0
PAN++ [35]	68.6	78.6	–	–	82.7	78.2	69.2	–	21.1
Boundary TextSpotter'22 [36]	66.2	78.4	46.1	73.0	82.5	77.4	71.7	–	13.4
SwinTextSpotter [17]	74.3	84.1	51.8	77.0	83.9	77.3	70.5	72.5	1.0
SRSTS [59]	78.8	86.3	–	–	85.6	81.7	74.5	–	18.7
TPSNet [56]	78.5	84.1	60.5	80.1	–	–	–	–	–
GLASS [46]	79.9	86.2	–	–	84.7	80.1	76.3	–	3.0
TESTR [68]	73.3	83.9	56.0	81.5	85.2	79.4	73.6	–	5.3
TTS [21]	78.2	86.3	–	–	85.2	81.7	77.4	–	–
ABINet++ [10]	77.6	84.5	60.2	80.3	84.1	80.4	75.4	76.5	10.6
DeepSolo [63]	79.7	87.0	64.2	81.4	86.8	81.9	76.9	–	17.0
ESTextSpotter (Ours)	**80.8**	**87.1**	**64.9**	**83.9**	**87.5**	**83.0**	**78.1**	**78.1**	4.3

Table 4: End-to-end text spotting results on VinText. ABCNet+D means adding the methods proposed in [41] to ABCNet. The same to Mask Textspotter v3+D.

Method	H-mean
ABCNet[33]	54.2
ABCNet+D[41]	57.4
Mask Textspotter v3[41]	53.4
Mask Textspotter v3+D[41]	68.5
SwinTextSpotter[17]	71.1
ESTextSpotter (Ours)	**73.6**

Table 5: End-to-end text spotting results and detection results on HUST-ART.

Method	Detection			E2E
	P	R	H	
DB[28]	95.31	74.62	83.71	–
DCLNet[3]	93.82	77.47	84.86	–
PAN++[54]	93.38	30.06	45.48	30.31
MaskTextSpotter v3[54]	88.31	80.82	84.40	71.23
ESTextSpotter (Ours)	**96.05**	**82.79**	**88.93**	**77.55**

Multilingual Text. We further evaluate ESTextSpotter using multilingual datasets. The results for the ReCTS can be found in Tables 1 and 3, which showcase ESTextSpotter's superior performance over the state-of-the-art method in both detection and text spotting. Notably, our method outperforms ABINet++, a method leveraging iterative language modeling for text spotting, by 1.6% in terms of text spotting performance. For VinText, the result is shown in Table 4, from which we can see ESTextSpotter outperforms the SwinTextSpotter by 2.5%. Note that ABCNet+D and Mask TextSpotter v3+D mean using the dictionary to train the model, which is not used by our method. For the HUST-ART [9], shown in Table 5, our method achieves the best performance on both detection and end-to-end recognition, significantly outperforming MaskTextSpotter v3. For the well-known multilingual benchmark ICDAR2019-MLT, the detection results and language-wise H-mean are shown in Table 2. Our method achieves the best performance in all languages except Bangla and Hindi. We provide some qualitative results in Figure 4.

4.3. Ablation Studies

We conduct ablation studies on the Total-Text to investigate the impact of different components in ESTextSpotter. For text spotting, the results contain a bias caused by randomness. Our experiments show that the pre-trained model had a bias of 0.2%, while the finetuned model had a bias of 0.5% in the end-to-end text spotting results. In the ablation studies, we use the pre-trained model to evaluate the results.

Comparison between implicit synergy and explicit synergy. To verify that our proposed explicit synergy achieves better synergy than the previous implicit synergy, we conduct experiments to validate that explicit synergy achieves better synergy compared to implicit synergy. The explicit synergy develops task-aware queries to conduct explicit interaction between detection and recognition within the decoder by modeling two distinct feature patterns for each task and interacting with each other. In contrast, the previous implicit synergy solely relies on shared features

Table 6: Ablation studies on Total-Text. "None" represents lexicon-free. IS means implicit interaction within the decoder. ES means explicit interaction within the decoder. VLC means the vision-language communication module in task-aware decoder. TAQI is the task-aware query initialization. TADN means the task-aware denoising training. REM means receptive enhancement module.

Method	IS	ES	TAQI	VLC	REM	TADN	Detection			End-to-End
							P	R	F	None
Baseline							89.5	84.8	87.1	68.2
Baseline	✓						87.0	81.4	84.1	70.2
Baseline		✓					90.6	85.0	87.7	70.4
Baseline		✓	✓				90.3	**86.5**	88.1	70.7
Baseline		✓	✓	✓			90.4	86.0	**88.3**	72.0
Baseline		✓	✓	✓	✓		90.3	86.2	88.2	72.7
ESTextSpotter		✓	✓	✓	✓	✓	**90.7**	85.3	87.9	**73.8**

(a) Total-Text (b) SCUT-CTW1500 (c) VinText

(d) ReCTS (e) ICDAR2015 (f) HUST-ART

Figure 4: Visualization results of our ESTextSpotter on different datasets. Best viewed in screen.

that overlook the divergent feature requirements of two tasks and lack explicit modules to ensure the interaction, resulting in limited synergy. The results are shown in Table 6. Although implicit synergy can improve text spotting performance, they often lead to a degradation in detection. In contrast, our proposed explicit synergy improves both detection and spotting performance, demonstrating its superior synergy.

Task-aware query initialization (TAQI) and vision-language communication module (VLC). The results shown in Table 6 demonstrate that the TAQI could lead to improvements of 0.4% and 0.3% for detection and end-to-end scene text spotting, respectively. Moreover, the VLC could further enhance the performance by 0.2% and 1.3% for detection and end-to-end scene text spotting, respectively. It demonstrates that conducting TAQI and VLC can promote stable explicit synergy and greatly enhance performance for text detection and spotting.

Receptive Enhancement Module (REM) and Task-aware denoising training (TADN). The results presented in Table 6 demonstrate that the REM can result in a 0.7% improvement in text spotting performance. Furthermore, TADN was employed, leading to an additional improvement of 1.1% in text spotting performance. Notably, TADN is more focused on text spotting rather than detection, as opposed to previous denoising training methods [23, 64].

5. Conclusion

In this paper, we present a simple yet effective Transformer with explicit synergy for text spotting. Previous implicit synergy can not fully realize the potential of two tasks. To address this issue, our approach explores explicit synergy to allow task-aware queries to explicitly model the discriminative and interactive features between text detection and recognition within a single decoder. Additionally, our proposed vision-language communication module enables task-aware queries to conduct interactions from a cross-modal perspective, thereby unleashing the potential of both text detection and recognition. Extensive experiments on a wide range of various benchmarks, including multi-oriented, arbitrarily-shaped, and multilingual datasets, consistently demonstrate that our method outperforms previous state-of-the-art approaches by significant margins. We hope our method can inspire further investigation on the explicit synergy in text spotting area.

Acknowledgement This research is supported in part by NSFC (Grant No.: 61936003) and Zhuhai Industry Core and Key Technology Research Project (no. 2220004002350).

References

[1] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9365–9374, 2019.

[2] Youngmin Baek, Seung Shin, Jeonghun Baek, Sungrae Park, Junyeop Lee, Daehyun Nam, and Hwalsuk Lee. Character region attention for text spotting. In *European Conference on Computer Vision*, pages 504–521. Springer, 2020.

[3] Yanguang Bi and Zhiqiang Hu. Disentangled contour learning for quadrilateral text detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 909–918, 2021.

[4] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[6] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. 2021.

[7] Chee-Kheng Ch'ng, Chee Seng Chan, and Cheng-Lin Liu. Total-text: toward orientation robustness in scene text detection. *International Journal on Document Analysis and Recognition (IJDAR)*, 23(1):31–52, 2020.

[8] Pengwen Dai, Sanyi Zhang, Hua Zhang, and Xiaochun Cao. Progressive contour regression for arbitrary-shape scene text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7393–7402, 2021.

[9] Wondimu DIKUBAB, Dingkang Liang, Minghui Liao, and Xiang Bai. Comprehensive benchmark datasets for amharic scene text detection and recognition. *Information Sciences*, 65(160106):1–160106, 2022.

[10] Shancheng Fang, Zhendong Mao, Hongtao Xie, Yuxin Wang, Chenggang Yan, and Yongdong Zhang. Abinet++: Autonomous, bidirectional and iterative language modeling for scene text spotting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[11] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7098–7107, 2021.

[12] Wei Feng, Wenhao He, Fei Yin, Xu-Yao Zhang, and Cheng-Lin Liu. Textdragon: An end-to-end framework for arbitrary shaped text spotting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9076–9085, 2019.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[14] Minghang He, Minghui Liao, Zhibo Yang, Humen Zhong, Jun Tang, Wenqing Cheng, Cong Yao, Yongpan Wang, and Xiang Bai. Most: A multi-oriented scene text detector with localization refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8813–8822, 2021.

[15] Tong He, Zhi Tian, Weilin Huang, Chunhua Shen, Yu Qiao, and Changming Sun. An end-to-end textspotter with explicit alignment and attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5020–5029, 2018.

[16] Jing Huang, Guan Pang, Rama Kovvuri, Mandy Toh, Kevin J Liang, Praveen Krishnan, Xi Yin, and Tal Hassner. A multiplexed network for end-to-end, multilingual ocr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4547–4557, 2021.

[17] Mingxin Huang, Yuliang Liu, Zhenghao Peng, Chongyu Liu, Dahua Lin, Shenggao Zhu, Nicholas Yuan, Kai Ding, and Lianwen Jin. Swintextspotter: Scene text spotting via better synergy between text detection and text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4593–4603, 2022.

[18] 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):1–20, 2016.

[19] 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 *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160. IEEE, 2015.

[20] 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 *2013 12th International Conference on Document Analysis and Recognition*, pages 1484–1493. IEEE, 2013.

[21] Yair Kittenplon, Inbal Lavi, Sharon Fogel, Yarin Bar, R Manmatha, and Pietro Perona. Towards weakly-supervised text spotting using a multi-task transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4604–4613, 2022.

[22] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[23] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022.

[24] Hui Li, Peng Wang, and Chunhua Shen. Towards end-to-end text spotting with convolutional recurrent neural net-

works. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5238–5246, 2017.

[25] M. Liao, P. Lyu, M. He, C. Yao, W. Wu, and X. Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):532–548, 2021.

[26] Minghui Liao, Guan Pang, Jing Huang, Tal Hassner, and Xiang Bai. Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 706–722. Springer, 2020.

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

[28] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11474–11481, 2020.

[29] Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[31] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022.

[32] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5676–5685, 2018.

[33] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9809–9818, 2020.

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

[35] 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 Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[36] Pu Lu, Hao Wang, Shenggao Zhu, Jing Wang, Xiang Bai, and Wenyu Liu. Boundary textspotter: Toward arbitrary-shaped scene text spotting. *IEEE Transactions on Image Processing*, 31:6200–6212, 2022.

[37] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67–83, 2018.

[38] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 20(11):3111–3122, 2018.

[39] Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowdhury, Dimosthenis Karatzas, Wafa Khlif, Jiri Matas, Umapada Pal, Jean-Christophe Burie, Cheng-lin Liu, et al. Icdar 2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1582–1587. IEEE, 2019.

[40] Nibal Nayef, Fei Yin, Imen Bizid, Hyunsoo Choi, Yuan Feng, Dimosthenis Karatzas, Zhenbo Luo, Umapada Pal, Christophe Rigaud, Joseph Chazalon, et al. Icdar 2017 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)*, volume 1, pages 1454–1459. IEEE, 2017.

[41] Nguyen Nguyen, Thu Nguyen, Vinh Tran, Triet Tran, Thanh Ngo, Thien Nguyen, and Minh Hoai. Dictionary-guided scene text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[42] Dezhi Peng, Xinyu Wang, Yuliang Liu, Jiaxin Zhang, Mingxin Huang, Songxuan Lai, Jing Li, Shenggao Zhu, Dahua Lin, Chunhua Shen, et al. Spts: single-point text spotting. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4272–4281, 2022.

[43] Liang Qiao, Ying Chen, Zhanzhan Cheng, Yunlu Xu, Yi Niu, Shiliang Pu, and Fei Wu. Mango: A mask attention guided one-stage scene text spotter. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, pages 2467–2476, 2021.

[44] Siyang Qin, Alessandro Bissacco, Michalis Raptis, Yasuhisa Fujii, and Ying Xiao. Towards unconstrained end-to-end text spotting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4704–4714, 2019.

[45] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

[46] Roi Ronen, Shahar Tsiper, Oron Anschel, Inbal Lavi, Amir Markovitz, and R Manmatha. Glass: Global to local attention for scene-text spotting. In *European Conference on Computer Vision*, pages 249–266. Springer, 2022.

[47] Guanglu Song, Yu Liu, and Xiaogang Wang. Revisiting the sibling head in object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11563–11572, 2020.

[48] 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 *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 4563–4572, 2022.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[50] Hao Wang, Pu Lu, Hui Zhang, Mingkun Yang, Xiang Bai, Yongchao Xu, Mengchao He, Yongpan Wang, and Wenyu Liu. All you need is boundary: Toward arbitrary-shaped text spotting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12160–12167, 2020.

[51] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *2011 International Conference on Computer Vision*, pages 1457–1464. IEEE, 2011.

[52] Pengfei Wang, Chengquan Zhang, Fei Qi, Shanshan Liu, Xiaoqiang Zhang, Pengyuan Lyu, Junyu Han, Jingtuo Liu, Errui Ding, and Guangming Shi. Pgnet: Real-time arbitrarily-shaped text spotting with point gathering network. *arXiv preprint arXiv:2104.05458*, 2021.

[53] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape robust text detection with progressive scale expansion network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9336–9345, 2019.

[54] Wenhai Wang, Enze Xie, Xiang Li, Xuebo Liu, Ding Liang, Yang Zhibo, Tong Lu, and Chunhua Shen. Pan++: Towards efficient and accurate end-to-end spotting of arbitrarily-shaped text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[55] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8440–8449, 2019.

[56] 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 *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5014–5025, 2022.

[57] Yuxin Wang, Hongtao Xie, Shancheng Fang, Jing Wang, Shenggao Zhu, and Yongdong Zhang. From two to one: A new scene text recognizer with visual language modeling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14194–14203, 2021.

[58] Yuxin Wang, Hongtao Xie, Zheng-Jun Zha, Mengting Xing, Zilong Fu, and Yongdong Zhang. Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11753–11762, 2020.

[59] Jingjing Wu, Pengyuan Lyu, Guangming Lu, Chengquan Zhang, Kun Yao, and Wenjie Pei. Decoupling recognition from detection: Single shot self-reliant scene text spotter. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1319–1328, 2022.

[60] Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, and Yun Fu. Rethinking classification and localization for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10186–10195, 2020.

[61] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1083–1090. IEEE, 2012.

[62] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[63] Maoyuan Ye, Jing Zhang, Shanshan Zhao, Juhua Liu, Tongliang Liu, Bo Du, and Dacheng Tao. Deepsolo: Let transformer decoder with explicit points solo for text spotting. *arXiv preprint arXiv:2211.10772*, 2022.

[64] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022.

[65] Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1577–1581. IEEE, 2019.

[66] Shi-Xue Zhang, Xiaobin Zhu, Jie-Bo Hou, Chang Liu, Chun Yang, Hongfa Wang, and Xu-Cheng Yin. Deep relational reasoning graph network for arbitrary shape text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9699–9708, 2020.

[67] Shi-Xue Zhang, Xiaobin Zhu, Chun Yang, Hongfa Wang, and Xu-Cheng Yin. Adaptive boundary proposal network for arbitrary shape text detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1305–1314, 2021.

[68] Xiang Zhang, Yongwen Su, Subarna Tripathi, and Zhuowen Tu. Text spotting transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9519–9528, 2022.

[69] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ICLR*, 2021.

[70] Yiqin Zhu, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, and Wayne Zhang. Fourier contour embedding for arbitrary-shaped text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3123–3131, 2021.