

# Supplementary Materials

Anonymous Authors

## 1 DETECTION RESULTS ON INVERSE-TEXT

In addition to comparing the recognition results of 'None' and 'Full', we also supplement the comparison of detection results here. We compare some mainstream end-to-end methods. DNTextSpotter employs the same data augmentation and the same additional pre-training datasets as DeepSolo, a mixture of Synth150K, MLT17, Total-Text, IC13, IC15, and TextOCR. After fine-tuning on the Total-Text for 2k iterations, DNTextSpotter directly applies these updated weights to assess performance on the InverseText dataset. It consistently outperforms current state-of-the-art methods, achieving 94.3% precision, 77.2% recall, and an F1-score of 84.9%.

**Table 1: Detection Performance on InverseText. The top two scores are shown in bold red and blue fonts.**

Method	Detection		
	Precision	Recall	F1
ABCNet [3](ResNet-50-FPN)	85.1	68.5	75.9
ABCNet v2 [4](ResNet-50-FPN)	87.1	64.6	74.2
TESTR [6](ResNet-50)	91.8	54.4	68.3
ESTextSpotter [1](ResNet-50)	78.7	71.4	74.9
DeepSolo [5](ResNet-50)	93.9	63.8	76.0
DeepSolo [5](ViTAEv2-S)	<b>95.1</b>	69.1	80.0
DNTextSpotter(ResNet-50)	94.3	<b>77.2</b>	<b>84.9</b>
DNTextSpotter(ViTAEv2-S)	<b>95.4</b>	<b>79.2</b>	<b>86.4</b>

## 2 DETAILS OF THE INSTABILITY MEASUREMENT

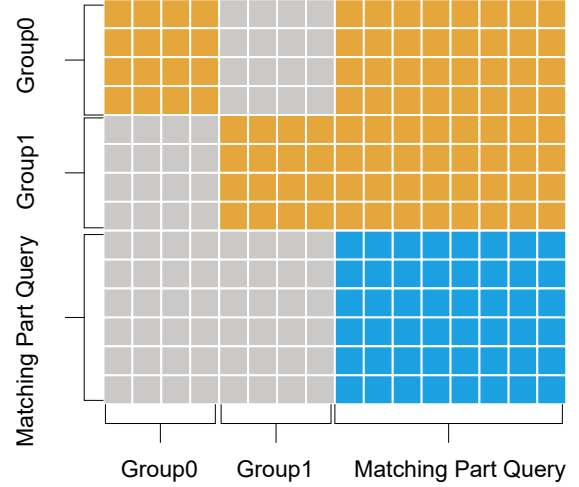
We analyze the instability of bipartite graph matching used by DN-DETR[2]. In the main text, we group every 10k iterations as one group. We adopt this setting here as well. For a training image, we represent the predicted text instances from transformer decoders at the  $i$ -th group as  $\mathbf{P}^i = \{P_0^i, P_1^i, \dots, P_{N-1}^i\}$ , where  $N$  signifies the total count of detected text instances, and the  $M$  ground truth text instances are denoted as  $\mathbf{G} = \{G_0, G_1, G_2, \dots, G_{M-1}\}$ . After bipartite matching, we generate a vector  $\mathbf{W}^i = \{W_0^i, W_1^i, \dots, W_{N-1}^i\}$  for the  $i$ -th iteration to capture the matching outcomes, defined by:

$$W_n^i = \begin{cases} m, & \text{if } P_n^i \text{ matches } G_m \\ -1, & \text{if } P_n^i \text{ matches nothing} \end{cases} \quad (1)$$

The stability for a single training image at iteration  $i$  is then determined by the variance between its  $W^i$  and  $W^{i+1}$ , calculated as:

$$IS^i = \sum_{k=0}^N \mathbb{I}(W_n^i \neq W_n^{i+1}) \quad (2)$$

Here,  $\mathbb{I}(\cdot)$  stands for the indicator function, where  $\mathbb{I}(z) = 1$  if  $z$  is true, and 0 otherwise. The overall stability for iteration  $i$  across the dataset is obtained by averaging these stability values for all images. Total Text contains a total of 1255 training images, with an



**Figure 1: We present an example of the attention mask when the number of the group is equal to 2. The values in the gray region are set to True to prevent information leakage from the denoising part to the matching part. The values in the orange and blue region are set to False, and the attention scores for this region are computed.**

average of 7.04 text instances per image, so the largest possible  $IS$  is  $7.04 \times 2 = 14.08$ . The  $IS$  visualization comparison results can be seen in the main text.

## 3 SINGLE ATTENTION MASK

We further present the attention mask in a graphical form to facilitate a better understanding for the readers. The attention mask  $\mathbf{A} = [a_{ij}]_{(g+2n) \times (g+2n)}$  is shown in the main text as follows:

$$a_{ij} = \begin{cases} 1, & \text{if } j < g \times 2n \text{ and } \lfloor \frac{i}{2n} \rfloor \neq \lfloor \frac{j}{2n} \rfloor; \\ 1, & \text{if } j < g \times 2n \text{ and } i \geq g \times 2n; \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

The visualization can be seen in Fig. 1.

## 4 MORE QUALITATIVE RESULTS ON BENCHMARKS

We provide more visualization results for the TotalText, CTW1500, ICDAR15, and InverseText datasets in Fig. 2, Fig. 3, Fig. 4, and Fig. 5. From these visualization results, it can be seen that we achieve advanced detection and recognition effects on texts of any shape. On inverse-like texts, our spotting performance also does not show any decline.

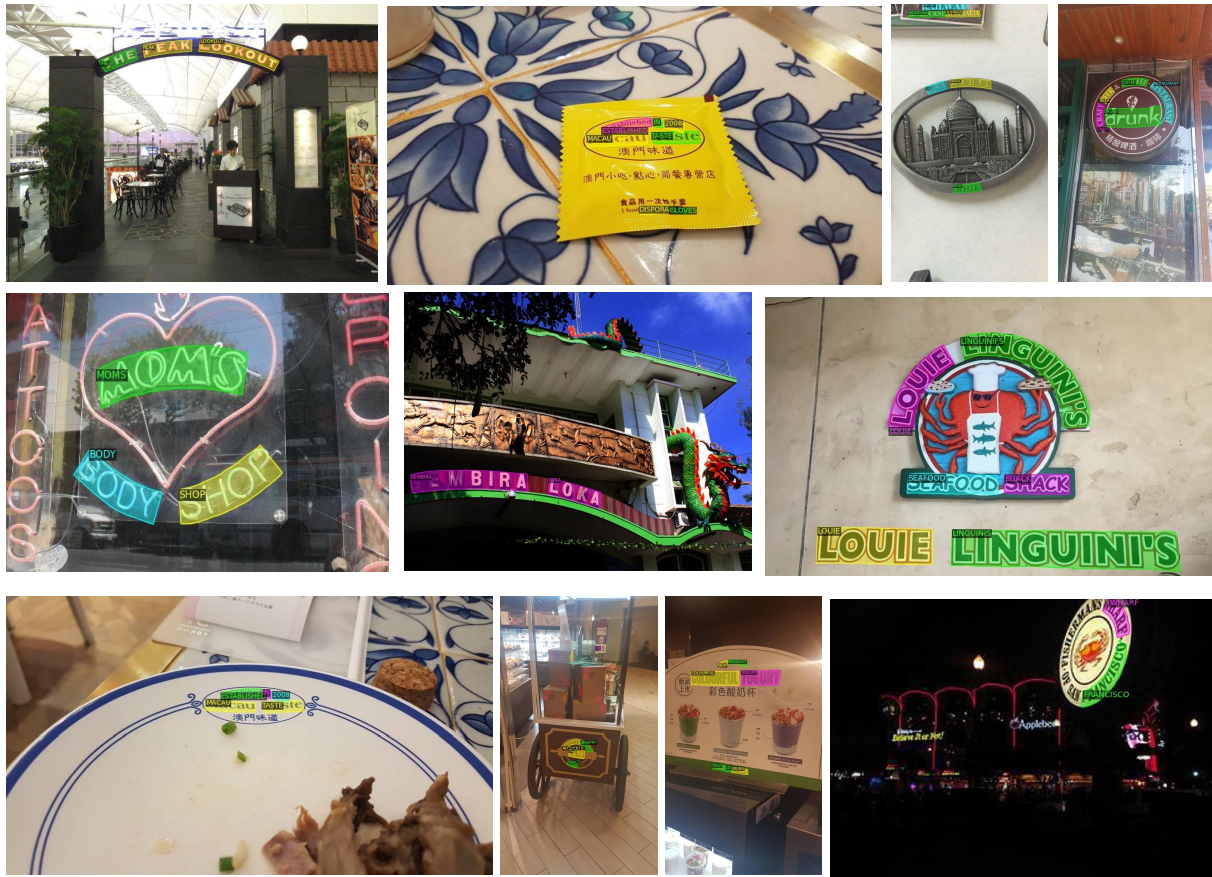


Figure 2: Qualitative results on Total Text.



Figure 3: Qualitative results on CTW1500.



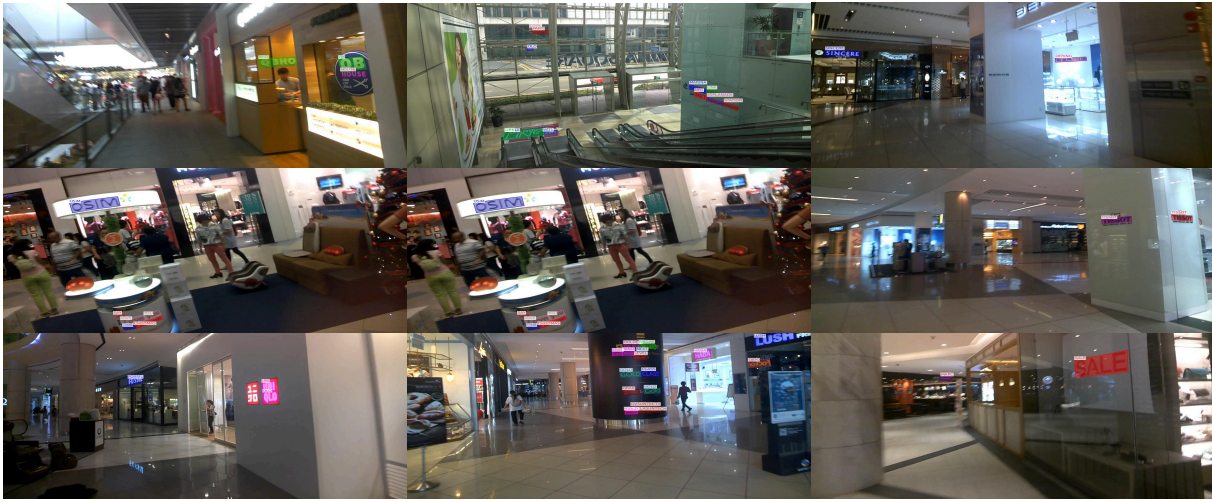


Figure 4: Qualitative results on ICDAR15.

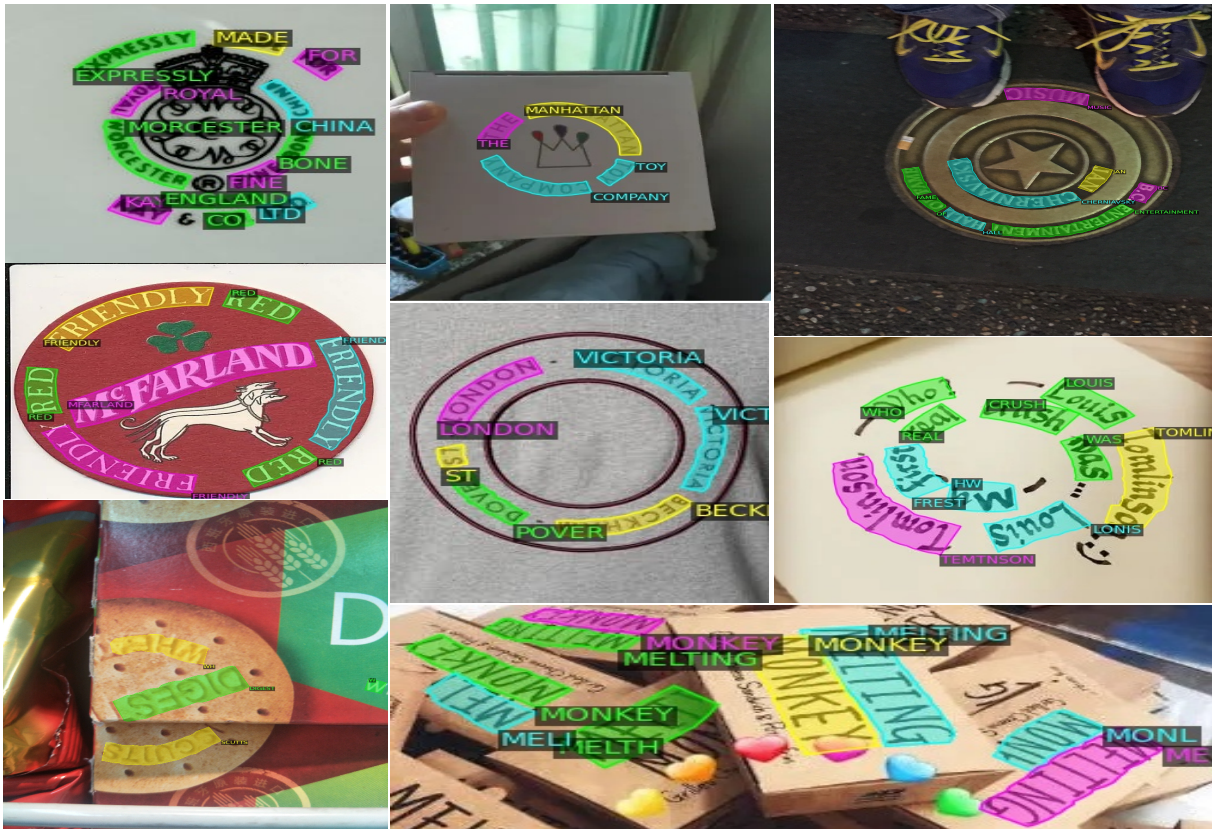


Figure 5: Qualitative results on InverseText.

## 5 LIMITATION AND DISCUSSION

Although DNTextSpotter achieves quite good performance, there are still some limitations. The most significant is the excessive overhead during training. Compared to the original vector shape of (bs,

100, 25, 256), during denoising training, the maximum shape can reach (bs, 200, 25, 256). Given that the computational complexity of the self-attention mechanism increases quadratically, the increased computational cost when the sequence length grows from 100 to

200 is non-negligible. DNTextSpotter was trained using 8 NVIDIA Tesla H800 GPUs, requiring approximately 26 hours of training time. Fortunately, the denoising training does not add any overhead during inference, making it a worthwhile method for actual deployment and application. Additionally, we have only applied the denoising training method to the evaluation of English scene text datasets and have not experimented with Chinese. We look forward to DNTextSpotter achieving similarly good results on Chinese datasets as well.

## REFERENCES

[1] Mingxin Huang, Jiaxin Zhang, Dezhi Peng, Hao Lu, Can Huang, Yuliang Liu, Xiang Bai, and Lianwen Jin. 2023. ETextSpotter: Towards Better Scene Text Spotting with Explicit Synergy in Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19495–19505.

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

[3] Yuliang Liu, Hao Chen, Chunhua Shen, Tong He, Lianwen Jin, and Liangwei Wang. 2020. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9809–9818.

[4] Yuliang Liu, Chunhua Shen, Lianwen Jin, Tong He, Peng Chen, Chongyu Liu, and Hao Chen. 2021. Abenet v2: Adaptive bezier-curve network for real-time end-to-end text spotting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 8048–8064.

[5] Maoyuan Ye, Jing Zhang, Shanshan Zhao, Juhua Liu, Tongliang Liu, Bo Du, and Dacheng Tao. 2023. Deepsolo: Let transformer decoder with explicit points solo for text spotting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19348–19357.

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