SemiETS: Integrating Spatial and Content Consistencies for Semi-Supervised End-to-end Text Spotting (Supplementary Material)

Table 1. The text detection results on Total-Text and ICDAR 2015 under the Partially Labeled Data setting. DeepSolo [6] is the baseline text spotter consistent with the main experiment.

Methods	Total-Text					ICDAR 2015				
	0.5%	1%	2%	5%	10%	0.5%	1%	2%	5%	10%
Supervised	77.1	80.2	81.4	81.8	83.6	75.5	74.9	76.7	81.4	82.6
STAC*	77.8	80.4	80.6	82.8	84.8	78.4	79.7	81.0	84.0	84.2
Mean-Teacher*	54.4	67.3	72.5	81.5	83.7	72.1	71.2	74.8	81.3	82.6
Soft Teacher*	59.2	61.8	73.6	78.9	81.3	70.2	73.4	75.4	80.9	80.2
UT v2*	56.0	67.3	76.7	81.4	84.5	69.3	69.2	73.3	79.0	80.3
Semi-DETR*	60.6	71.8	74.8	79.8	82.0	78.9	79.6	82.2	83.6	84.4
SemiETS (Ours)	**78.8**	**80.8**	**82.7**	**84.5**	**85.4**	**80.2**	**82.9**	**83.4**	**85.8**	**86.1**

Table 2. Results of additional domain adaptation experiment (IC15 → Total-Text; Total-Text → TextOCR).

Methods	D_l	D_u	Det-F1	None	Full	D_l	D_u	Det-F1	None
Supervised	-	-	44.6	34.9	40.6	-	-	32.3	22.5
Supervised	IC15	-	72.9	65.0	75.8	TT	-	54.6	41.2
STAC*	IC15	TT	73.8	67.3	75.3	TT	TextOCR	53.2	37.0
Mean-Teacher*	IC15	TT	76.1	69.0	77.9	TT	TextOCR	52.6	33.0
Soft Teacher*	IC15	TT	68.4	64.4	71.6	TT	TextOCR	47.4	33.0
UT v2*	IC15	TT	72.0	68.0	75.3	TT	TextOCR	48.6	26.1
Semi-DETR*	IC15	TT	61.9	55.9	63.5	TT	TextOCR	38.1	30.8
SemiETS	IC15	TT	**78.6**	**71.5**	**80.0**	TT	TextOCR	**55.3**	**43.4**

Table 3. Comparison to using VLLMs as zero-shot text spotters or label generators using 2% labeled data on Total-Text.

Settings	Methods	Det-F1	None	Full
Zero-shot	InternVL2-8B	0.3	0.0	0.1
	Qwen2-VL-7B	1.8	0.6	1.4
Label Generator	InternVL2-8B	0.0	0.0	0.0
	Qwen2-VL-7B	1.2	1.1	1.2
	SemiETS	**82.7**	**73.4**	**82.2**

Table 4. Ablation study on the training stages of applying MMS using 2% labeled data setting.

Settings	Applied stages		Detection			E2E	
	O2M	O2O	P	R	F1	None	Full
w/o MMS			94.1	71.6	81.3	72.4	80.4
Full	✓	✓	95.5	66.5	78.4	72.6	79.2
O2O		✓	93.5	74.2	**82.7**	**73.4**	**82.2**

A. Additional Experimental Results

A.1. Text Detection Results

As shown in Tab. 1, the proposed SemiETS achieves state-of-the-art text detection results on arbitrary-shaped and multi-oriented scene text under all proportions. Nevertheless, the performance of several existing semi-supervised object detection (SSOD) methods even declines, especially in low data proportions. We attribute this to two aspects. Firstly, the irregular shape of texts increases the difficulty of detection. Secondly, the accumulated error caused by noisy pseudo labels disturbs the optimization. SemiETS reduces noisy pseudo labels using progressive sample assignment and explicitly enhances the complementarity of detection and recognition by mutual mining, thereby facilitating the performance of both tasks.

A.2. Additional Domain Adaptation Results

To simulate diverse domain shifts, We add domain adaptation settings, *i.e.*, from IC15 to Total-Text and from Total-Text to TextOCR. Results in Tab. 2 further demonstrate the consistent improvements in domain adaptation of SemiETS.

A.3. Comparison to VLLMs

Since generalist vision-language large models (VLLMs) have shown promising performance on various tasks recently, we select recent representative open-source VLLMs, *i.e.,* InternVL2 [1] and Qwen2-VL [5], to verify their effectiveness on our task. However, results in Fig. 3 reveals their limitations in text spotting. Firstly, as competitive baselines, their spotting results are unsatisfactory. Secondly, we use them as pseudo-label generators to generate pseudo labels on unlabeled data and then train spotters. Results are even worse as low-quality labels dominate the optimization to the false direction. It is because VLLMs are good at understanding tasks but are unsuitable for fine-grained perception tasks, indicating the value of our work in the era of VLLMs.

Table 5. Comparison of different additional data for Total-Text.

Settings	Det-F1	None	Full
Supervised	87.3	79.7	87.0
+ MSRA-TD500	86.6	80.2	86.8
+ COCOText	87.3	80.2	87.4
+ TextOCR	**87.5**	**81.7**	**87.6**

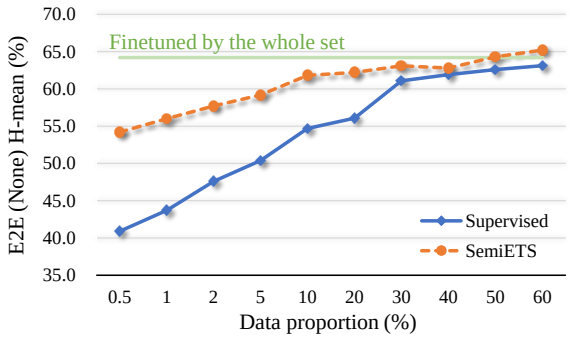


Figure 1. The E2E (None) performance trend of ABCNet on Total-Text under the Partially Labeled Data setting. The green indicates the model finetuned using the whole annotated training set [4].

B. Extensive Ablation Experiments

Training stages. For DETR-based spotters, we introduce the stage-wise hybrid matching strategy [7] to the assignment of PSA to boost the training efficiency, dividing the training process into one-to-many (O2M) and one-to-one (O2O) stage. As shown in Tab 4, applying the Mutual Mining Strategy (MMS) only during the O2O stage achieves the best detection and text spotting results. However, introducing MMS into the O2M stage would cause a decrease in detection performance due to the restriction of recall. In early iterations, the pseudo labels generated by the teacher are usually sparse and less reliable. While exploring the potentially high-quality positive proposals using the O2M matching, low-quality predictions would be introduced simultaneously, which might mislead the focus of MMS. Therefore, MMS is applied only to the O2O stage to refine the guidance after adequate high-quality proposals can be generated.

Diversity of additional data. We further explore various unlabeled data sources in the Fully Labeled Data setting on Total-Text in Tab. 5. Improvements demonstrate the robustness of SemiETS to utilize unlabeled data. In particular, higher quality and diversity help handle more complex scenes and text styles, bringing more performance gain.

Parameter study. We study the influence of hyper-parameters in Tab. 6. We empirically choose $\mathcal{T}_R = 0.7$ and $\lambda = 20$ by default.

Table 6. Parameter study.

(a) The threshold $\mathcal{T}_R$.

$\mathcal{T}_R$	0.5	0.6	0.7	0.8	0.9
Det-F1	82.0	82.6	**82.7**	82.5	81.8
E2E (None)	72.4	72.6	**73.4**	72.9	73.2
E2E (Full)	81.0	81.3	**82.2**	81.4	81.0

(b) The scale factor λ.

λ	1	10	20	50	100
Det-F1	81.7	82.3	**82.7**	82.3	82.5
E2E (None)	73.0	73.1	**73.4**	73.1	73.1
E2E (Full)	81.6	81.2	**82.2**	81.4	81.5

Figure 2. Visualization of pseudo labels generated by SemiETS in typical scenarios.

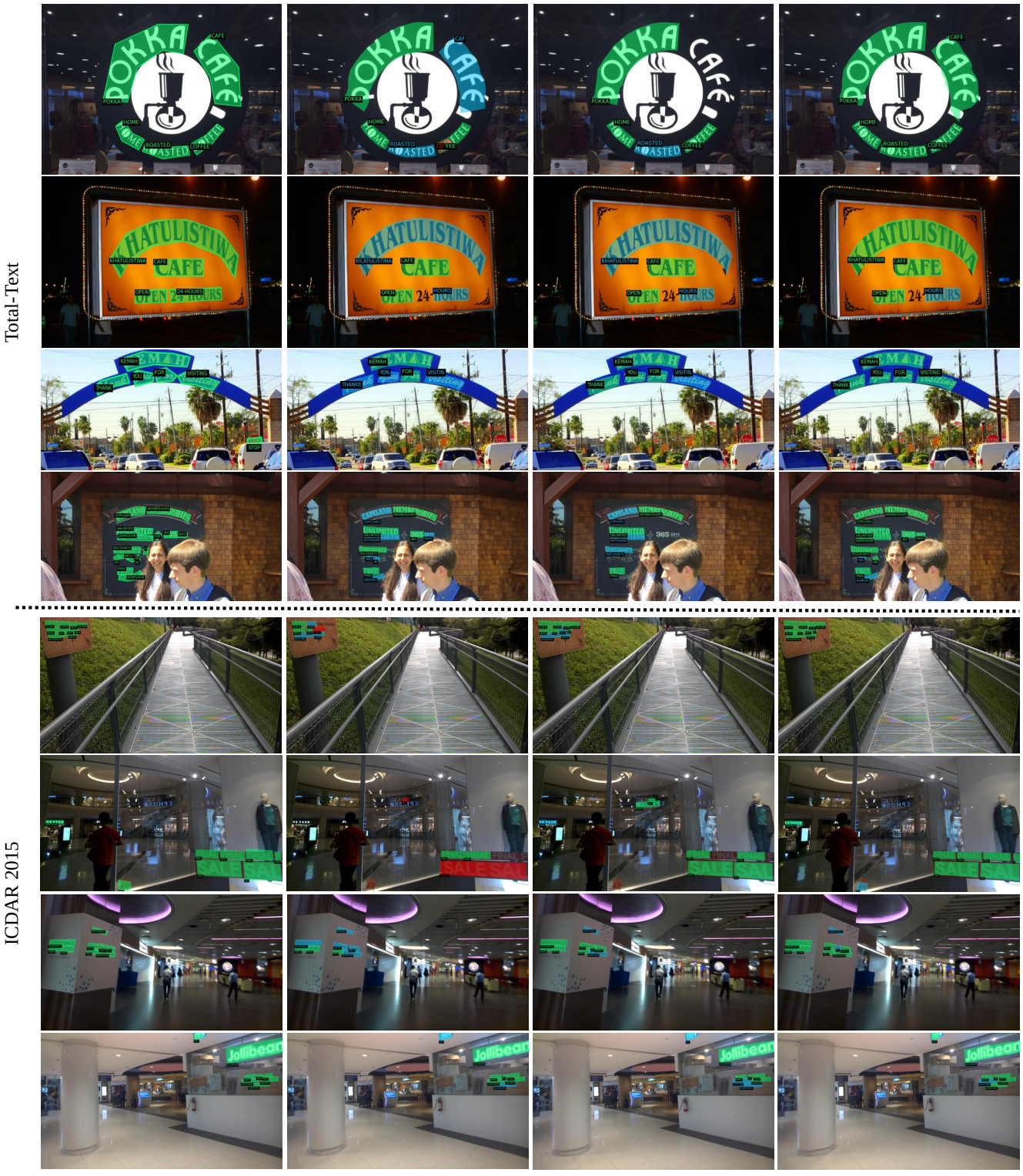
C. Performance Trend

We gradually increase the proportion of labeled data of Total-Text under the Partially Labeled Data setting and display the performance trend of ABCNet [4] on E2E H-mean without lexicon in Fig. 1. SemiETS can significantly boost text spotting performance compared to the supervised baseline, and the improvement is more notable when using less labeled data. Furthermore, as the proportion of annotated data increases, E2E H-mean continues growing. When only using 50% labeled data, SemiETS even outperforms the model finetuned using the whole labeled training set of Total-Text referred from [4], demonstrating the potential of the proposed framework to effectively reduce labeling cost and explore useful information from unlabeled data.

D. More Visualization Results

D.1. Pseudo Labels

We visualize pseudo labels generated by SemiETS in several challenging scenarios shown in Fig. 2 to examine its effectiveness and potential limitations. 1) Arbitrary-shaped texts increase the difficulty of obtaining precise localization labels. SemiETS can handle them with the proposed MMS to rectify text location. 2) Complex text fonts would lead to incorrect pseudo recognition labels. SemiETS can distinguish them and alleviate noisy recognition labels while still making use of reliable localization labels with the proposed PSA. 3) Dense texts would lead to label omission or shift due to adjacent interference. SemiETS exhibits decent pseudo label generation ability to some extent, as it imposes fine-grained constraints. However, for some extremely tiny and blurry texts, SemiETS still faces challenges.



(a) Ground truth (b) Supervised baseline (c) Semi-DETR (d) SemiETS

Figure 3. Qualitative results on Total-Text and ICDAR 2015. True positives are indicated in green. Text instances in blue are localized accurately but recognized incorrectly. Instances in red are inaccurately localized.

D.2. Qualitative Results

We visualize representative qualitative results from Total-Text [2] and ICDAR 2015 [3] in Fig. 3. SemiETS demonstrates superior performance in detecting and localizing curved and multi-oriented scene texts while significantly minimizing recognition errors. This improvement stems from its progressive sample assignment mechanism, effectively mitigating noisy supervision signals for text recognition, and its mutual mining strategy, which aims at extracting important guidance information. The robustness of SemiETS gets further validated in challenging scenarios, including incidental and densely distributed scene texts.

References

[1] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 24185–24198, 2024. 1

[2] Chee-Kheng Chng, Chee Seng Chan, and Cheng-Lin Liu. Total-text: toward orientation robustness in scene text detection. *Int. J. Document Anal. Recognit.*, 23(1):31–52, 2020. 4

[3] 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 *Int. Conf. Doc. Anal. Recognit.*, pages 1156–1160. IEEE, 2015. 4

[4] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9806–9815, 2020. 2

[5] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[6] 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. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19348–19357, 2023. 1

[7] Jiacheng Zhang, Xiangru Lin, Wei Zhang, Kuo Wang, Xiao Tan, Junyu Han, Errui Ding, Jingdong Wang, and Guanbin Li. Semi-detr: Semi-supervised object detection with detection transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 23809–23818, 2023. 2