Multi-scenario Overlapping Text Segmentation with Depth Awareness

Supplementary Material

A. Visualizations

To verify the effectiveness of our method, we visualized the inference outputs from both the baseline and our method in Fig. 9. The results reveal that, at low overlap rates (as seen in rows 1 and 2), the overlap relationships are not immediately apparent, but our method successfully captures these subtle relationships. Conversely, at high overlap rates (illustrated in rows 3, 4, and 5), we observe that complete stroke-level overlap can lead the model to incorrectly conclude that no overlap exists, as evidenced by the output from Mask2Former. In contrast, our method effectively delineates the overlapping areas by leveraging depth information, allowing us to accurately identify and separate these overlaps.

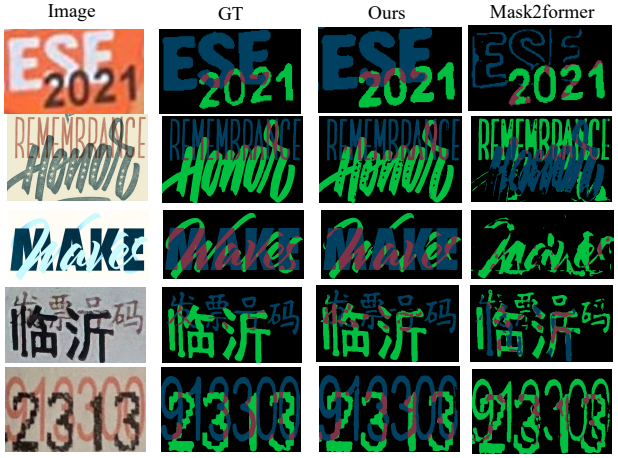

Figure 9. Visualizations of the qualitative comparison between the baseline and our method.

B. Sensitivity to depth maps

We acknowledge the critical role of accurate depth estimation for resolving overlapping text. To systematically evaluate the impact of depth map quality on model performance, we conducted dedicated experiments. Recognizing that pretrained depth estimators achieve high accuracy on natural scene images but face challenges on document images, we stratified our evaluation dataset into distinct Scene and Document scenarios. As detailed in Tab. 8, using Mask2Former as the baseline with identical configurations, our experiments demonstrate:

1. Strong Scene Generalization: Pre-trained depth estimators generalize effectively to scene text images. The structural priors learned from natural scenes enable reliable depth prediction for overlapping text instances, which our depth-guided decoder successfully leverages to enhance segmentation performance.

2. Robustness in Challenging Cases: Regarding concerns about performance degradation due to poor depth estimates, we specifically evaluated challenging Document scenarios where depth estimation is inherently difficult (e.g., flat-layout documents with minimal texture/cues). In these cases, while the depth guidance did not yield significant performance gains beyond the baseline, the model achieved segmentation performance comparable to the baseline (Mask2Former without depth guidance). This indicates that the depth guidance mechanism does not detrimentally impact performance when depth cues are unreliable.

Methods	Doc		Scene	
	IoU_{Ov}	$mIoU_{Text}$	IoU_{Ov}	$mIoU_{Text}$
Baseline	61.61	77.99	51.25	53.47
Ours	**62.00**	**78.30**	**55.46**	**58.26**

Table 8. Ablation study in doc. (hard to estimate depth) and scene (easy to estimate depth) scenarios.