**SwinTextSpotter: Scene Text Spotting via Better Synergy between Text Detection and Text Recognition (Supplementary Material)**

Mingxin Huang†, Yuliang Liu†, Zhenghao Peng, Chongyu Liu, Dahua Lin
Shenggao Zhu, Nicholas Yuan, Kai Ding, Lianwen Jin

1South China University of Technology 2Chinese University of Hong Kong 3Huawei Cloud AI
4IntSig Information Co., Ltd 5Peng Cheng Laboratory

eelwjin@scut.edu.cn

1. Qualitative Comparisons

We make some qualitative analysis with previous method which is shown in Fig 1. It can be seen that previous methods failed on the difficult text instance such as “Party”, while SwinTextSpotter can handle such case by exploiting the synergy of text detection and recognition.

Intuitively, the detection result of SwinTextSpotter is more accurate.

![Figure 1. Qualitative analysis of SwinTextSpotter and other existing methods. Best view in screen.](image)

2. Ablation study of Recognition Conversion

We verify the effectiveness of other components without RC which helps to better reveal the effectiveness of RC.

![Table 1. Ablation study on Recognition Conversion.](image)

<table>
<thead>
<tr>
<th>Method</th>
<th>Total-Text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Det-Hmean</td>
</tr>
<tr>
<td>SwinTextSpotter-withou RC</td>
<td>82.8</td>
</tr>
<tr>
<td>SwinTextSpotter</td>
<td>83.2</td>
</tr>
</tbody>
</table>

From Table 1, the performance, that using other components without RC, drops from 83.2% to 82.8% for detection and 66.9% to 63.4% for end-to-end scene text spotting.

3. Comparison different backbone in different frameworks

We also try to replace ResNet50 with Swin-Transformer on ABCNet [3]. From Table 2, the result can be improved by 1.8% with Swin-Transformer in text spotting. But there is no improvement for detection. It is similar to the case in SwinTextSpotter.

![Table 2. Comparison different backbone on different architectures on Total-Text. ABC-R50 means ABCNet with ResNet50. ABC-Swin means ABCNet with SwinTransformer. Det. means detection result. E2E means end-to-end text spotting result.](image)

<table>
<thead>
<tr>
<th>ABC-R50</th>
<th>ABC-Swin</th>
<th>Our-R50</th>
<th>Our-Swin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Det.</td>
<td>E2E</td>
<td>Det.</td>
<td>E2E</td>
</tr>
<tr>
<td>86.0</td>
<td>67.1</td>
<td>86.0</td>
<td>69.8</td>
</tr>
<tr>
<td>86.0</td>
<td>72.4</td>
<td>87.3</td>
<td>74.0</td>
</tr>
</tbody>
</table>

References


[2] Minghui Liao, Guan Fang, Jing Huang, Tal Hassner, and Xiaang Bai. Mask textspotter v3: Segmentation proposal network

