Supplementary Materials-
Attention Where It Matters: Rethinking Visual Document Understanding with Selective Region Concentration

This supplementary material presents a comparative case study of SeRum with end-to-end methods and OCR-dependent methods for handling challenging images. The evaluation utilizes several test sets, such as SROIE [2], CORD [5], Ticket [1], and DocVQA [4].

Comparison results with end-to-end methods. SeRum is compared with Donut [3], the current state-of-the-art end-to-end document understanding method, which decoding text directly from image features. However, Donut suffers from generating overly long results, leading to instability, and attention mechanism deviation and confusion. In contrast, SeRum excels at decoding the form of localized visual tokens of interest, leading to significant improvements in both of these drawbacks.

As illustrated in Figure 1.(a) to (c), Donut generates an abnormal sequence of text due to the interference of redundant characters, and it cannot correctly parse all the key information. In contrast, SeRum possesses the ability to identify the key area of interest and perform decoding process in isolation. Additionally, as shown in Figure 1.(d) to (f), Donut exhibits a tendency to misinterpret the location of text, whereas SeRum is capable of correctly identifying the text and its location within the image. Overall, SeRum demonstrates a superior performance relative to Donut.

Comparison results with OCR-dependent methods. This section evaluates the performance of SeRum on handwritten or blurry text images. Handwritten text recognition poses a significant challenge to OCR systems due to the inherent complexity and variability of handwritten fonts. Handwritten characters exhibit a high degree of variation in shape, size, slant, *etc.* as shown in Figure 2.(a) to (g). Besides, the stability of the system can also be significantly impacted by the presence of blurry text, as exemplified in Figure 2.(h) to (o).

The SeRum model simplifies the character recognition pipeline by integrating all stages into a single model, achieving end-to-end optimization that improves accuracy and reduces error propagation. Additionally, the model utilizes attention mechanisms to extract robust features from input images and effectively utilize contextual information.

Our findings indicate that the SeRum method can synthesize the context and help improve the OCR recognition

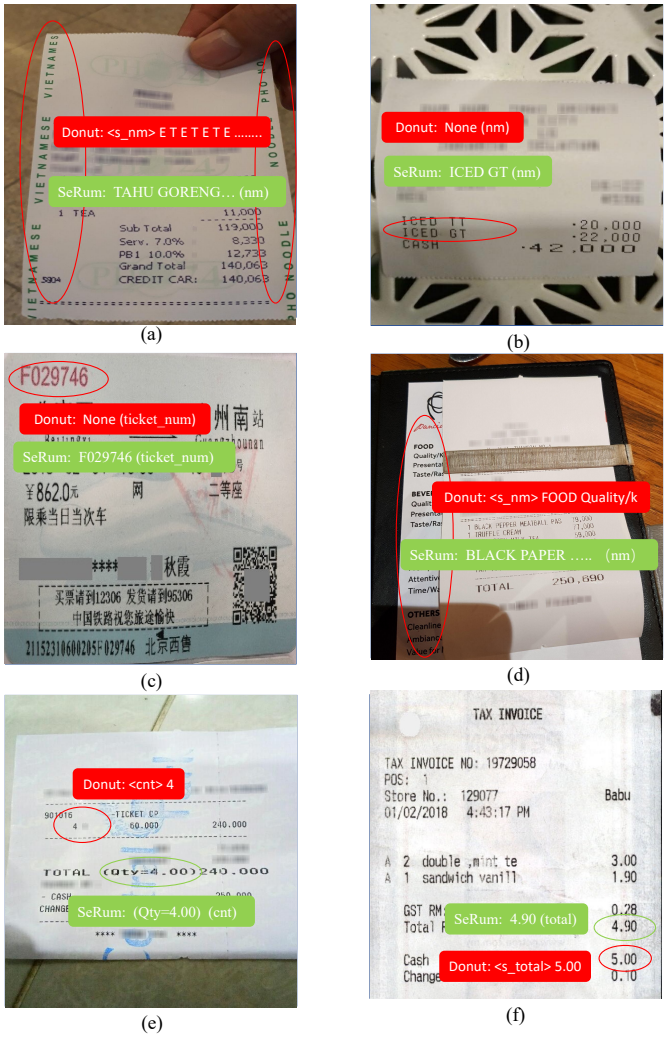

Figure 1. Challenging cases encountered in our study, including instances of redundant text on the border, superimposed images, *etc.* The red and green boxes represent the outputs of Donut and SeRum, respectively. Best viewed in color.

results. For example, in Figure 2.(d), the SeRum model identifies the word 'home' after a phone number, indicating that it is a home phone rather than a less common term like 'hame'. As shown in Figure 2.(n), the SeRum model distinguished 'MART' from 'MAPT', despite the visual similarity of the latter due to vagueness.

Figure 2. Examples of challenging cases, such as handwritten, blurred, and missing characters, illustrating the difficulties faced by OCR systems. These characters are known to pose difficulties in accurate recognition. The red and green boxes represent the outputs of OCR engine and SeRum, respectively. Best viewed in color.

References

[1] He Guo, Xiameng Qin, Jiaming Liu, Junyu Han, Jingtuo Liu, and Errui Ding. EATEN: entity-aware attention for single shot visual text extraction. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 254–259. IEEE, 2019.

[2] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. ICDAR2019 competition on scanned receipt OCR and information extraction. *CoRR*, abs/2103.10213, 2021.

[3] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, volume 13688 of *Lecture Notes in Computer Science*, pages 498–517. Springer, 2022.

[4] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE, 2021.

[5] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: A consolidated receipt dataset for post-ocr parsing. 2019.