Supplementary of Multi-scale Cell-based Layout Representation for Document Understanding

1. Ablation study

To compare with different methods we use the same hyper parameters in every comparison experiment, which results in some experiments results are worse than the results in the main paper. We use the F1 score of named entity recognition to evaluate the methods.

Table 1. **Comparison results with and without the proposed data augmentation** Ours(P) means our proposed spatial position representation. LMv3 means the LayoutLMv3 model.

Method	Data Augmentation	F1 (%)
LMv3$_{BASE}$	No	21.41
LMv3$_{BASE}$	Yes	22.50
Ours(P)	No	25.56
Ours(P)	Yes	**26.17**

Data augmentation on FUNSD. We use the proposed data augmentation on the FUNSD [2] dataset to evaluate the performance of data augmentation on form documents. The results are shown in Table 1. The F1 score is improved by 1.09% on LayoutLMv3$_{BASE}$ [1] and 0.61% on Our(P) by using the proposed data augmentation, respectively. Therefore, the proposed data augmentation can be used not only for receipts, but also for form documents.

Table 2. **Comparison results using LMv3$_{LARGE}$ on the FUNSD dataset.** Word means the number of word-level cells and Token is the number of token-level cells in the input data. P means applying the proposed spatial position representation to the baseline model. M means applying the multi-scale layout to the baseline model. LM means the LayoutLMv3 model.

Method	Word	Token	Scale factor Θ	F1 (%)
LMv3$_{LARGE}$	512	0	-	22.52
Ours(P)	512	0	-	26.15
Ours(M)	300	212	-	59.63
Ours(P+M)	300	212	-	**62.61**
Ours(P)	512	0	0.1	26.92
Ours(P)	512	0	0.2	**27.06**
Ours(P)	512	0	0.3	25.86

LayoutLMv3$_{LARGE}$ on FUNSD. Since we did not show many results using LayoutLMv3$_{LARGE}$, we apply the proposed spatial position representation, multi-scale layout and proposed data augmentation method to LayoutLMv3$_{LARGE}$, and use F1 score of the FUNSD dataset to evaluate them. As a result, three methods improved the performance of LayoutLMv3$_{LARGE}$, as presented in Table 2. Different from the results on the CORD dataset, an appropriate scale factor on the FUNSD dataset is 0.2 which is less than the CORD dataset. The named entities cover larger areas in the documents from the FUNSD dataset than from the CORD dataset. Because the number of named entities of each document image in the FUNSD dataset is more than in the CORD dataset. Using a big scale factor Θ would make the bounding boxes overlap each other. Therefore, the appropriate scale factor Θ of the proposed data augmentation of the FUNSD dataset is smaller than that of the CORD dataset.

Table 3. **Comparison results using LMv3$_{BASE}$ on the CORD dataset.** Pre presents using pre-trained baseline model. Word shows the number of word-level cells in the input data. Token shows the number of token-level cells in the input data. W means using the classification results of word-level cells to calculate F1 score. T means using the classification results of token-level cells to calculate F1 score. P means applying the proposed spatial position representation to the baseline model. M means applying the multi-scale layout to the baseline model. LM means the LayoutLMv3 model.

Method	Pre	Word	Token	Metric	F1 (%)
Ours(P)	No	512	0	W	60.68
Ours(P)	No	300	0	W	57.86
Ours(M)	No	300	212	T	84.51
Ours(P+M)	No	300	212	T	85.25
LMv3$_{BASE}$	Yes	512	0	W	96.56
Ours(P)	Yes	512	0	W	96.97
Ours(P)	Yes	300	0	W	96.52
Ours(M)	Yes	300	212	T	97.01
Ours(M+P)	Yes	300	212	T	**97.23**
Ours(M+P)	Yes	300	212	W	97.12

LayoutLMv3$_{BASE}$ on CORD. We use LayoutLMv3$_{BASE}$ as a baseline model to test the multi-scale cell-based layout on the CORD dataset [3]. The results are shown in Table 3. To insert the token-level cell into the input data and do not modify the structure of the baseline model, we have to reduce the number of word-level cells in the input data. We product experiments to check the influence of reducing word-level cells. The F1 score decreases with the reduction in the number of word cells in the input data. Therefore, we consider that improvement of F1 score is the result of utilizing the multi-scale layout.

Evaluation on LayoutLM and LayoutLMv2. To evaluate the generic of the cell-based layout, we apply the pro-

Table 4. **Evaluation the cell information using LayoutLM and LayoutLMv2.** LMv2 means LayoutLMv2, and LM means LayoutLM. P means applying the proposed spatial position representation to the baseline model.

Model	Pre-train	F1 (%)	Acc. (%)
LM_{BASE}	No	18.91	42.94
$\text{Ours(P)}_{LM_{BASE}}$	No	20.09	44.38
LMv2_{BASE}	No	20.55	45.87
$\text{Ours(P)}_{LMv2_{BASE}}$	No	23.18	47.48
LMv2_{BASE}	Yes	82.76	-
$\text{Ours(P)}_{LMv2_{BASE}}$	Yes	83.09	-

posed spatial position representation to LayoutLM_{BASE} [4] and LayoutLMv2_{BASE} [5] using the FUNSD dataset. We use the row and column index to replace the x and y coordinates of the bottom right corner for each bounding box and do not change the structure of the original models. Noted that we do not use image embedding in LayoutLM, therefore the modality of the LayoutLM is $Text + Layout$. As shown in Table 4, the cell information could improve the performance of all baseline models. We consider the cell information could provide useful information for layout representation and improve the performance of existing methods.

References

[1] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. *arXiv preprint arXiv:2204.08387*, 2022. 1

[2] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *International Conference on Document Analysis and Recognition Workshops (ICDARW)*, pages 1–6, 2019. 1

[3] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019. 1

[4] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. pages 1192–1200, 2020. 2

[5] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *ACL*, pages 2579–2591, 2021. 2