# Visual vs Textual Features: In the battleground of instance document layout Segmentation.

***Objective of the thesis:*** *Instance document layout segmentation (DLS) is the task of identifying the different layout elements such as text, images, tables, and graphs, in a document image. One of the critical decisions in this task is the choice of features used to represent the layout elements. In this thesis, we compare the performance of visual with textual features and multi-modal (visual + textual) features in instance document layout segmentation.*

***A brief Literature Review:*** *In the past few decades several architectures have been proposed to solve this problem. As convolutional neural networks (CNN) deal with spatial features and make use of the same knowledge across all image locations, many researchers [1, 2, 3] make use of CNNs to solve DLS. Though CNNs work efficiently well for small objects, they fail for large object segmentation. On the other hand, transformers based networks [4, 5, 6] segment the large objects properly but don't converge for the small objects. To mitigate this trade off people start using textual information [7, 8, 9] as in documents, text contains important information to classify a document object. In order to use the advantage of both the modality, multimodal instance document layout segmentation is a current trend of research [10, 11, 12]. But text mining is more difficult than visual features extraction and needs large computational resources. For multi-modal analysis we need a large scale pre training dataset which isn't publicly available. Also, we need a good OCR to get the text and its corresponding bounding boxes but they are very expensive.*

***Proposed Thesis Statement:*** *In order to solve this problem, we have applied a different learning paradigm to enrich the visual features which helps to improve the document layout analysis performance. First, we use a domain adaptation strategy to make use of pretrained weights of completely different domains (e.g. MS-COCO). This has been performed by hybrid bipartite matching. Next in order to correctly segment the small objects we utilize a swin transformer with contrastive denoising training paradigm. It can outperform the large scale dataset which have sufficient data points to train but fail to perform for small scale dataset with large variability.*

*To solve the problems of small scale datasets, we introduce a few shot learning paradigm. From the previous work, we already know which objects are affecting the performance of the system. So in this setting, we first train the system with only those classes for 100K epochs and test on the rest of the classes. Then we finetune the system for 3K epoch with the rest of the classes and test on the whole test set. This proposed setting is called Decouple Boosting in few shot instance segmentation.*

*However, all the above approaches use positional embeddings. But relational embeddings are also an important piece of information to explore. In order to do that, we applied a self-supervised learning paradigm to understand the attention of the network during segmentation. We treat those attention regions as nodes and build a fully connected graph with* *those nodes and use graph sage to perform the instance segmentation with the help of these graphs.*

***Conclusion:*** *Layout analysis is an important part of document understanding as the performance of Key information extraction, Visual Question Answering and other tasks depend on the document layout analysis. So, we need a fast and efficient system for this task. With visual features we can save time in feature extraction, feature mining as well as resource cost. This excites us to explore different learning paradigms which can help to advance the current state-of-the-art.*

***Reference***

*[1] Prusty, A., Aitha, S., Trivedi, A. and Sarvadevabhatla, R.K., 2019, September. Indiscapes: Instance segmentation networks for layout parsing of historical indic manuscripts. In 2019 International Conference on Document Analysis and Recognition (ICDAR) (pp. 999-1006). IEEE.*

*[2] Liu, Y., Si, C., Jin, K., Shen, T. and Hu, M., 2020. FCENet: an instance segmentation model for extracting figures and captions from material documents. IEEE Access, 9, pp.551-564.*

*[3] Zhao, P., Wang, W., Cai, Z., Zhang, G. and Lu, Y., 2021. Accurate fine-grained layout analysis for the historical Tibetan document based on the instance segmentation. IEEE Access, 9, pp.154435-154447.*

*[4] Biswas, S., Banerjee, A., Lladós, J. and Pal, U., 2022. DocSegTr: an instance-level end-to-end document image segmentation transformer. arXiv preprint arXiv:2201.11438.*

*[5] Cheng, H., Jian, C., Wu, S. and Jin, L., 2022, November. SCUT-CAB: A New Benchmark Dataset of Ancient Chinese Books with Complex Layouts for Document Layout Analysis. In Frontiers in Handwriting Recognition: 18th International Conference, ICFHR 2022, Hyderabad, India, December 4–7, 2022, Proceedings (pp. 436-451). Cham: Springer International Publishing.*

*[6] Kim, H., Choi, J., Park, S. and Jung, Y., 2022. Layout Aware Semantic Element Extraction for Sustainable Science & Technology Decision Support. Sustainability, 14(5), p.2802.*

*[7] Martínek, J., Lenc, L. and Král, P., 2020. Building an efficient OCR system for historical documents with little training data. Neural Computing and Applications, 32, pp.17209-17227.*

*[8] Zhu, W., Sokhandan, N., Yang, G., Martin, S. and Sathyanarayana, S., 2022, June. DocBed: A multi-stage OCR solution for documents with complex layouts. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 11, pp. 12643-12649).*

*[9] Qiao, L., Jiang, H., Chen, Y., Li, C., Li, P., Li, Z., Zou, B., Guo, D., Xu, Y., Xu, Y. and Cheng, Z., 2022, October. DavarOCR: A Toolbox for OCR and Multi-Modal Document Understanding. In Proceedings of the 30th ACM International Conference on Multimedia (pp. 7355-7358).*

*[10] Appalaraju, S., Jasani, B., Kota, B.U., Xie, Y. and Manmatha, R., 2021. Docformer: End-to-end transformer for document understanding. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 993-1003).*

*[11] Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W. and Zhang, M., 2020. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. arXiv preprint arXiv:2012.14740.*

*[12] Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C. and Wei, F., 2022, October. Dit: Self-supervised pre-training for document image transformer. In Proceedings of the 30th ACM International Conference on Multimedia (pp. 3530-3539).*